ALEXANDRU I. ȘCHIOP
MIHAI STANCU

MATEMATICĂ APLICATĂ ÎN TEHNICA DE CALCUL

XII

ALEXANDRU I. ȘCHIOP • MIHAI STANCU

MATEMATICĂ APLICATĂ ÎN TEHNICA DE CALCUL

XII

Clasa a XII-a

Manual pentru licee cu profil
de matematică-fizică și informatică

EDITURA DIDACTICĂ ȘI PEDAGOGICĂ, BUCUREȘTI

Capitolul I

ERORI. CONCEPTE DE BAZĂ ÎN CALCULUL NUMERIC

§ 1. Introducere

Calculul aproximativ joacă un rol central în matematica aplicată.

În practică ne mulțumim foarte adesea cu o valoare aproximativă a cantității pe „care o găsim", alteori ignorăm unele cantități mici în comparație cu altele, iar nu de puține ori „datele de intrare", ale problemei puse, sînt inexacte.

Introducem următoarele notații (nu sînt exacte din punct de vedere matematic) utile în practică:

a $\ll b$ (sau $b \gg a$) citim: „a este mult mai mic decît b" (sau „b este mult mai mare decît a").

Evident, acest „mult mai mare" sau „mult mai mic" depinde de contextul problemei $\left(\text{uneori } a \ll b \text{ dacă } a < \frac{1}{4} \cdot b\text{, în timp ce în alte situații } a \ll b\right.$ dacă $\left. a < \frac{1}{500} \cdot b\right)$.

$a \approx b$; „a este aproximativ egal cu b" și înseamnă că $|a - b| \leq c$, unde c este ales de contextul problemei.

§ 2. Surse de erori

Orice rezultat al unui calcul numeric poate fi influențat (uneori dezastruos!) de diferite tipuri de erori.

a) Erori în datele de intrare. De exemplu dacă coeficienții unui sistem liniar sînt fracții ordinare, la transformarea lor în fracții zecimale se introduc erori $\left(\frac{1}{3} = 0{,}3333333\right)$. Alt exemplu, considerarea în calculul cu numere iraționale (π, $\sqrt{2}$) a unui număr finit de zecimale.

Uneori datele de intrare sînt rezultatele unor măsurători care sînt sistematic afectate de erori.

În problemele tehnico-științifice (rezolvarea cu calculatorul a ecuațiilor diferențiale, de exemplu) apar foarte des sisteme liniare algebrice în care

coeficienții sînt integrale ce trebuie calculate aproximativ și deci coeficienții sistemului liniar sînt afectați de erori.

b) Erori de rotunjire în timpul calculului. Dacă facem un calcul utilizînd numai t cifre în toate numerele cu care lucrăm, atunci produsul exact a două numere este format cu $2t$ sau $2t - 1$ cifre. Dar noi nu putem utiliza la pasul următor de calcul decît rotunjirea la t cifre. Astfel se produc erori care pe parcursul unui lung proces de calcul pot afecta serios rezultatele.

c) Erori de trunchiere. Acestea sînt erori ce provin din limitarea la o „parte" din termenii unui proces de calcul înainte de a ajunge la limită. De exemplu, din suma progresiei geometrice infinite $1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \dots$ reținînd numai primii patru termeni, diferența dintre suma progresiei și acest rezultat se numește eroare de trunchiere.

Tot o eroare de trunchiere facem și atunci cînd aproximăm derivata unei funcții $f(x)$ într-un punct x_0 prin raportul:

$$\frac{f(x_0 + h) - f(x_0)}{h},$$

unde h este „mic".

De asemenea același tip de eroare se face cînd pe o porțiune mică în jurul unui punct aproximăm o funcție cu tangenta la graficul său în punctul respectiv.

d) Erori de simplificare a modelului și erori comise de om și mașină.

Uneori, în practică, în unele modele, idealizăm anumiți parametri pentru a putea prelucra modelul.

De asemenea se poate greși în programul pentru calculator sau în perforarea cartelelor, erori de operare etc.

§ 3. Erori relative și erori absolute

Fie $\widetilde{a}$ o valoare aproximativă pentru cantitatea a (a este valoarea exactă) avem:

Eroarea absolută în $\widetilde{a}$ este $\widetilde{a} - a$.

Eroarea relativă în $\widetilde{a}$ este $(\widetilde{a} - a)/a$ dacă $a \neq 0$.

Foarte des eroarea relativă se dă în procente. De exemplu, 4% înseamnă o eroare relativă de 0,04.

Acum „$a - \widetilde{a}$" este **corecția** ce trebuie adăugată lui $\widetilde{a}$ pentru a obține valoarea exactă.

Corecția și eroarea au aceeași valoare absolută dar semne contrarii.

Eroarea poate fi pozitivă sau negativă dar noi putem introduce o margine pentru eroare care este întotdeauna un număr pozitiv.

Dacă $a = \widetilde{a} \pm \varepsilon$ înseamnă că $|\widetilde{a} - a| \leqslant \varepsilon$

De exemplu $a = 0,462 \pm 0,011$ înseamnă:

$$0,451 \leqslant a \leqslant 0,473$$

§ 4. Sisteme de numerație. Reprezentarea numerelor în virgulă mobilă

În general sîntem obișnuiți să gîndim și să calculăm cu numere în baza 10. Astfel numărul 257,424 reprezintă:

$$2 \cdot 10^2 + 5 \cdot 10^1 + 7 \cdot 10^0 + 4 \cdot 10^{-1} + 2 \cdot 10^{-2} + 4 \cdot 10^{-3}$$

Orice număr real în baza 10 are o reprezentare unică în forma de mai sus. Se exceptează cazul în care, în scrierea pozițională a unui număr zecimal (în baza 10) cifra 9 apare de o infinitate de ori.

Exemple:

1) *numărul* 99,9999 ... *reprezintă același număr ca* 100
2) *numărul* 2,219999 ... *reprezintă același număr ca* 2,22.

În general pentru reprezentarea unui număr real putem utiliza o bază oarecare B (număr natural) cu $B \geqslant 2$ și atunci se poate arăta (cu excepții ca mai sus) că orice număr real pozitiv admite o reprezentare unică de forma:

$$a_n \cdot B^n + a_{n-1} \cdot B^{n-1} + \dots + a_1 \cdot B + a_0 + a_{-1} \cdot B^{-1} + a_{-2} \cdot B^{-2} + \dots$$

unde coeficienții „a" se numesc **cifre în sistemul cu baza** B și sînt numere întregi pozitive cu proprietatea că:

$$0 \leqslant a_i \leqslant B - 1.$$

Avantajul mare al scrierii unui număr într-o bază care conduce la o scriere pozițională de forma:

$$a_n a_{n-1} \dots a_1 a_0,\ a_{-1}\, a_{-2} \dots$$

unde $a_n a_{n-1} \dots a_1 a_0$ reprezintă *partea întreagă* iar $0,\ a_{-1}\, a_{-2} \dots$ *partea fracționară* a numărului real, este că putem da reguli simple pentru operațiile aritmetice.

Cu cît baza este un număr natural mai mic cu atît aceste reguli devin mai simple. Acest argument stă la baza faptului că cele mai multe calculatoare operează în baza 2 (în sistem binar cu două cifre 0 și 1).

Amintim regulile de adunare și înmulțire în sistemul binar:

$$0 + 0 = 0 \qquad 0 \cdot 0 = 0$$
$$0 + 1 = 1 \qquad 0 \cdot 1 = 0$$
$$1 + 0 = 1 \qquad 1 \cdot 0 = 0$$
$$1 + 1 = 10 \qquad 1 \cdot 1 = 1.$$

Alte sisteme de numerație utilizate sînt cele **octale** cu cifre de la 0 la 7 și **hexadecimale** cu cifre de la 0 la 9 și literele A, B, C, D, E, F de la **zece** la **cincisprezece.**

De exemplu:

$$(13{,}25)_{10} = (1101{,}01)_2 = (15{,}2)_8 = (\mathrm{D}{,}4)_{16}$$

unde prin $()_{10}$, $()_2$, $()_8$, $()_{16}$ *am notat baza de lucru.*

Virgulele utilizate aici separă partea întreagă de partea fracționară și se numesc respectiv: **virgulă zecimală, virgulă binară, virgulă octală, virgulă hexadecimală.**

Un calculator este echipat de regulă cu două tipuri de operații aritmetice, anume **calculul cu virgulă fixă** și **calculul cu virgulă mobilă. Virgula este zecimală dacă baza este** *10*, **binară dacă baza este** *2* etc.

În cele ce urmează ne vom ocupa în special de baza zecimală.

Calculul cu virgula mobilă zecimală se efectuează cu un număr constant de cifre după cum se va observa în cele ce urmează.

Ne oprim acum mai îndeaproape la calculul cu numere reprezentate în virgulă mobilă.

Numim reprezentare zecimală normalizată cu virgulă mobilă a numărului a **o reprezentare de forma:**

$$a = m \cdot 10^q, \text{ unde } 0{,}1 \leqslant |m| < 1$$

m **este număr real pozitiv sau negativ,**
q **este număr întreg pozitiv sau negativ.**

O astfel de reprezentare este totdeauna posibilă, pentru $a \neq 0$.

Variabila m poartă numele de **mantisă** și este o parte fracționară, iar q se numește **exponent** și este întreg.

De exemplu *numărul* 542,36 *va fi reprezentat ca*:
$0{,}54236 \cdot 10^3$, deci 0,54326 *este mantisa iar* 3 *este exponentul.*

Alte exemple:

1) $4527 = 0{,}4527 \cdot 10^4$ mantisa este 0,4527 și exponentul 4
2) $0{,}0002617 = 0{,}2617 \cdot 10^{-3}$ mantisa este 0,2617 și exponentul -3
3) $-42{,}15 = -0{,}4215 \cdot 10^2$ mantisa este $-0{,}4215$ și exponentul 2
4) $-0{,}012 = -0{,}12 \cdot 10^{-1}$ mantisa este $-0{,}12$ și exponentul -1

Un număr zecimal în virgulă mobilă este memorat, în multe calculatoare, cu mantisa și exponentul în aceeași celulă de memorie.

Celula de memorie însă nu are prevăzută decît o singură poziție binară pentru semn și astfel se ridică problema memorării cînd și mantisa și exponentul sînt negativi. Soluția este în a aduna exponentului o constantă pozitivă de care se ține seamă în calculele aritmetice.

De exemplu, *într-o mașină zecimală cu două cifre pentru exponent vom aduna 50 la toți exponenții (ceea ce înseamnă că exponenții din exemplele de mai sus vor fi reprezentați în interiorul calculatorului ca* 54, 47, 52, 49).

Pe două poziții zecimale rezervate exponentului putem memora 100 numere diferite (numerele de la 0 la 99). Cum exponentul poate lua și valori negative vom considera un interval de valori de lungime 100 simetric în jurul originii:

$$[-50, \; +49]$$

Deci exponentul real va lua valori în intervalul [— 50, + 49] însă în calculator va fi memorat (pe cele două poziții rezervate lui) un număr cu 50 mai mare.

Într-un calculator numărul de cifre pentru q și m este limitat.

Să studiem acum operațiile în virgulă mobilă.

1) *Să presupunem că dorim să adunăm numerele:*

$x_1 = 112,5$ *în virgulă mobilă memorat ca* $0,1125 \cdot 10^3$

$x_2 = 15$ *în virgulă mobilă memorat ca* $0,15 \cdot 10^2$

Așa cum sînt reprezentate numerele nu putem aduna mantisele deoarece virgula numerelor nu este aliniată.

Diferența dintre cei doi exponenți este 1, vom deplasa atunci mantisa celui de al doilea număr (numărul cu exponentul mai mic) spre dreapta cu un loc, adică scriem:

$$x_2 = 0,015 \cdot 10^3$$

și acum putem aduna:

$$x_1 + x_2 = (0,1125 + 0,015) \cdot 10^3 = 0,1275 \cdot 10^3$$

Observăm că am lucrat aici cu o mantisă formată cu 4 cifre.

2) *Să considerăm un alt exemplu:*

$$x_1 + x_2 = 0,1213 \cdot 10^1 + 0,4691 \cdot 10^{-1}.$$

Deplasăm din nou mantisa numărului cu exponent mai mic spre dreapta cu un număr de poziții egal cu diferența exponenților, deci în acest caz cu 2 și obținem:

$$x_1 + x_2 = 0,1213 \cdot 10^1 + 0,004691 \cdot 10^1 = 0,1259 \cdot 10^1.$$

Se observă că am pierdut cifrele 9 și 1 ale mantisei deplasate în afara capacității de calcul (deoarece am considerat mantisa ca avînd patru cifre).

3) *Presupunem acum că avem de efectuat următoarea adunare:*

$$0,9917 \cdot 10^3 + 0,1230 \cdot 10^3.$$

Numerele de adunat au același ordin de mărime (același exponent) deci singura operație ce urmează a se efectua este adunarea mantiselor:

$$0,9917 + 0,1230 = 1,1147.$$

Rezultatul adunării mantiselor este un număr mai mare ca 1.

Noi am definit însă mantisa ca o fracție zecimală subunitară,

$$0,1 \leqslant |m| < 1.$$

Operația care urmează a se efectua este **deplasarea mantisei spre dreapta cu o poziție zecimală și adunarea la exponent a unei unități.** Această operație se numește **renormalizare.**

Deci:

$$0{,}9917 \cdot 10^3 + 0{,}1230 \cdot 10^3 = 0{,}1114 \cdot 10^4$$

(din nou prin deplasare spre dreapta am pierdut cifra 7 deoarece calculul se efectuează cu 4 zecimale).

4) *Fie acum de efectuat produsul a două numere x_1 și x_2 în virgulă mobilă:*

$$0{,}1763 \cdot 10^{-2} \times 0{,}2036 \cdot 10^5.$$

Normal, se înmulțesc mantisele,

$$m_1 \cdot m_2 = 0{,}1763 \times 0{,}2036 = 0{,}0358,$$

iar exponenții se adună $q_1 + q_2 = 3$.

Deci,

$$x_1 \cdot x_2 = 0{,}0358 \cdot 10^3.$$

Nici de data aceasta mantisa (în modul) nu este cuprinsă în intervalul [0,1 ; 1).

Operația de renormalizare constă în acest caz în deplasarea spre stînga a mantisei 0,0358 **cu o poziție zecimală și scăderea unei unități din exponent** (adică $x_1 \cdot x_2 = 0{,}3580 \cdot 10^2$).

Observăm deci că operația de renormalizare, în urma efectuării unei operații aritmetice constă în „aducerea" mantisei în intervalul [0,1 ; 1) cu „ajustarea" corespunzătoare a exponentului.

> *Observație*. În majoritatea calculatoarelor înmulțirea se efectuează pe registru de memorie dublu (adică o celulă dublă de memorie). Dacă se înmulțesc două numere de lungime „t", rezultatul de „$2t$" sau „$2t - 1$" cifre este obținut în acest registru de lungime dublă. Dacă după efectuarea înmulțirii este necesară renormalizarea (numai cu deplasarea spre stînga în acest caz) cifra de pe ultima poziție (poziția „t") a numărului va proveni din a doua parte a registrului dublu de memorie și va fi deci corectă. În exemplul de mai sus $x_1 \cdot x_2 = 0{,}03588468 \cdot 10^3$ în registrul de lungime dublă iar rezultatul renormalizat pe patru zecimale va fi $0{,}3588 \cdot 10^2$.

Dăm în figura I.1 organigrama adunării a două numere în virgulă mobilă.

Amintim că, în organigrama din figura 1, un dreptunghi reprezintă un bloc de calcul și de atribuire (se efectuează operațiile din membrul drept al semnului egal și valoarea obținută se atribuie parametrului din partea stîngă a semnului egal) iar un romb reprezintă un bloc de comparație și de decizie.

Presupunem acum că avem un calculator în care sînt prevăzute 5 poziții zecimale pentru mantisă și două poziții zecimale pentru exponent.

Cel mai mare număr normalizat ce poate fi memorat cu aceste facilități va fi:

$0{,}99999 \cdot 10^{99}$ în realitate $0{,}99999 \cdot 10^{49}$ deoarece am convenit să adunăm exponentului real numărul 50.

Iar cel mai mic număr normalizat (în modul) va fi:

$$0{,}10000 \cdot 10^0 \text{ în realitate } 0{,}10000 \cdot 10^{-50}$$

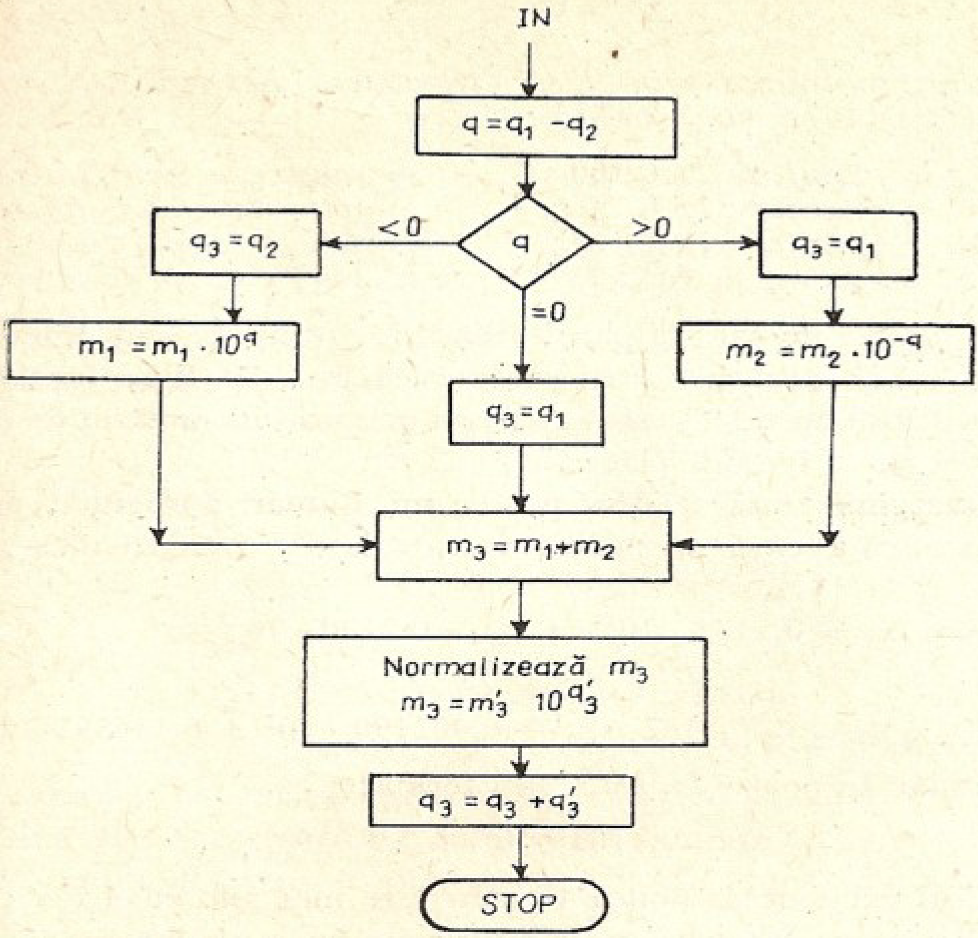


Fig. I. 1. Organigrama adunării a două numere în virgulă mobilă.

$$(x_1 = m_1 \cdot B^{q_1},\ x_2 = m_2 \cdot B^{q_2},\ x_1 + x_2 = m_3 \cdot B^{q_3}).$$

Cînd printr-o operație aritmetică între doi operanzi am depășit una sau alta din limitele arătate mai sus (un număr mai mare de $1{,}0 \cdot 10^{49}$ — (depășire flotantă superioară) sau un număr mai mic de $1{,}0 \cdot 10^{-49}$ — (depășire flotantă inferioară)) calculul se oprește.

Exemple:

1) $0{,}12641 \cdot 10^{-43} \times 0{,}98763 \cdot 10^{-24}$
2) $0{,}45631 \cdot 10^{42} \times 0{,}18764 \cdot 10^{15}$

vor cauza oprirea procesului de calcul.

§ 5. Rotunjirea și rotunjirea prin tăiere (trunchiere)

Există două căi de rotunjire a numerelor la un număr de zecimale („t") fixat anume, rotunjirea prin eliminarea tuturor zecimalelor care se găsesc la dreapta poziției t (**rotunjire prin tăiere**) și a doua cale este **rotunjirea obișnuită** adică printre numerele ce se pot forma cu „t" zecimale alegem numărul cel mai apropiat de numărul dat.

Exemple

Rotunjirea și rotunjirea prin tăiere (trunchiere) la trei zecimale în numerele de mai jos se face după cum urmează:

1) 0,1597 *se rotunjește la* 0,160; *se rotunjește prin tăiere la* 0,159
2) 0,2456 *se rotunjește la* 0,246; *se rotunjește prin tăiere la* 0,245
3) 0,1574 *se rotunjește la* 0,157; *se rotunjește prin tăiere la* 0,157
4) 0,23652 *se rotunjește la* 0,237; *se rotunjește prin tăiere la* 0,236

Un mare număr de calculatoare utilizează rotunjirea prin tăiere în rezultatele obținute după fiecare operație aritmetică, însă dat fiind faptul că numărul de cifre utilizate în calcul este mult mai mare decît numărul de cifre în datele problemei nu se ivesc neplăceri.

Aflarea marginii erorii relative pentru un număr aproximat prin tăiere. Să presupunem că avem la dispoziție un calculator cu 4 cifre disponibile pentru mantisă.

Să adunăm $x_1 = 0,1142 \cdot 10^1$ cu $x_2 = 0,2190 \cdot 10^{-1}$.

Obținem:

$$x_3 = x_1 + x_2 = 0,1142 \cdot 10^1 + 0,002190 \cdot 10^1 = 0,11639 \cdot 10^1.$$

Acest număr se poate exprima sub forma:

$$x_3 = 0,1163 \cdot 10^1 + 0,9 \cdot 10^{-3}.$$

Exponentul celui de al doilea termen este mai mic cu 4 decît exponentul primului termen. Întotdeauna cînd lucrăm cu o mașină cu patru cifre în mantisă, exponentul celui de al doilea termen trebuie să fie mai mic cu patru decît exponentul primului termen.

Deci dacă avem un calculator cu „t" cifre în mantisa unui număr reprezentat în virgulă mobilă atunci orice rezultat al unui calcul (obținut prin aplicarea oricărei din cele patru operații aritmetice) se poate reprezenta (evident înainte de rotunjire) ca:

$$b = m_b \cdot 10^q + m'_b \cdot 10^{q-t}$$

unde m_b este un număr format cu „t" cifre.

Valorile posibile pe care le poate lua m_b sînt:

$$0,1 \leqslant |m_b| < 1$$

în timp ce:

$$0 \leqslant |m'_b| < 1.$$

Spunem că facem o aproximație prin tăiere (trunchiere) dacă, pur și simplu nu-l luăm în considerare pe m'_b și acesta este cazul practic cel mai des întîlnit la calculatoare.

Eroarea relativă maximă se obține cînd m'_b *este cel mai mare și* m'_b *cel mai mic.*

Valoarea maximă posibilă a lui m'_b *este mai mică decît* 1 *iar valoarea minimă a lui* m_b *este* 0,1.

Deci limita maximă a erorii relative va fi:

$$\frac{1 \cdot 10^{q-t}}{0,1 \cdot 10^{q}} = 10^{-t+1}.$$

La rotunjire, prin același procedeu, obținem o eroare relativă dată de

$$\frac{0,5 \cdot 10^{q-t}}{0,1 \cdot 10^{q}} = 0,5 \cdot 10^{-t+1}.$$

În general pentru o bază arbitrară B avem adevărat următorul criteriu:
Presupunem că lucrăm în baza B cu „t" cifre în mantisă (cifra binară care ne dă semnul numărului nu este socotită). Atunci *orice număr real în virgulă mobilă poate fi reprezentat cu o eroare relativă care nu depășește* **unitatea mașinii definită** de:

$$u = \begin{cases} \frac{1}{2} B^{1-t} & \text{dacă utilizăm rotunjirea} \\ B^{1-t} & \text{dacă utilizăm tăierea.} \end{cases}$$

În cele mai multe calculatoare „u" se află între 10^{-6} și 10^{-15}.

Să presupunem acum că operanzii cu care lucrăm ar fi exact reprezentați în mașină. Nu este sigur că rezultatul va fi exact reprezentat (să ne gîndim doar la faptul că rezultatul înmulțirii a două numere de lungime „t" are „$2t$" cifre).

Vom nota cu $fl(x+y)$, $fl(x-y)$, $fl(x \cdot y)$, $fl(x/y)$ rezultatele operațiilor de adunare, scădere, înmulțire și împărțire în virgulă mobilă pe care mașina le memorează în urma efectuării acestor operații și aproximării prin tăiere sau rotunjire.

Conform criteriului precedent asupra erorii relative avem

$$\frac{fl(x \odot y) - (x \odot y)}{(x \odot y)} \leqslant u$$

unde $\odot$ este una din operațiile $+$, $-$, $\cdot$, $/$

Să răspundem acum la întrebarea: *a) Care este unitatea mașinii pentru un calcul cu virgulă mobilă cu 5 cifre zecimale pentru mantisă?*

$$u = \begin{cases} 0,5 \cdot 10^{-4} & \text{dacă se utilizează rotunjirea} \\ 1 \cdot 10^{-4} & \text{dacă se utilizează tăierea.} \end{cases}$$

b) Fie acum $x_1 = 0,12347 \cdot 10^{3}$ un număr rezultat din calcul.
Care este marginea pentru eroarea absolută în acest calcul dacă s-a utilizat **rotunjirea obișnuită**?

Eroarea absolută e_1 este mărginită de $e_1 = 1 \cdot 10^{3} \cdot 0,5 \cdot 10^{-5} = 0,5 \cdot 10^{-2}$.

c) *Pentru numerele rezultate din calcul* $x_2 = 0,12347 \cdot 10^{18}$ și $x_3 = 0,20006 \cdot 10^{10}$ *erorile absolute vor fi mărginite de*: $e_2 = 0,5 \cdot 10^{13}$ și $e_3 = 0,5 \cdot 10^{5}$.

d) Care este marginea pentru eroarea absolută în calculul celor trei numere de mai sus dacă s-a folosit **rotunjirea prin tăiere**?

$e_1 = 0,1 \cdot 10^{-1}$

$e_2 = 0,1 \cdot 10^{14}$

$e_3 = 0,1 \cdot 10^{6}$

Putem cere variabilei x_1 o precizie absolută de $0,1 \cdot 10^{-6}$? Dar o precizie de $0,1 \cdot 10^{-40}$? Evident că nu, deoarece valorile de $0,1 \cdot 10^{-6}$ și $0,1 \cdot 10^{-40}$ sînt mult prea mici în comparație cu eroarea cu care calculatorul ne „furnizează" acest număr $(0,5 \cdot 10^{-2})$.

> *Observație*. La efectuarea unui calcul cu virgulă mobilă trebuie să avem permanent în vedere că ultima cifră zecimală a mantisei este afectată de erori de rotunjire sau trunchiere. Deci dacă am obținut dintr-un calcul cu virgula mobilă, cu 5 cifre zecimale pentru mantisă, numărul $x = 0,23483 \cdot 10^7$ marginea pentru eroarea absolută este de ordinul $0,000005 \cdot 10^7 = 0,5 \cdot 10^2$ în cazul cînd se aplică rotunjirea sau $0,00001 \cdot 10^7 = 0,1 \cdot 10^3$ cînd se aplică truchierea și deci nu putem cere o precizie mai bună decît $0,5 \times 10^2$ respectiv $0,1 \cdot 10^3$.

Acum vom arăta că **asociativitatea în adunarea în virgulă mobilă nu se păstrează.**

Presupunem că lucrăm cu o mașină cu șapte cifre în mantisă. Fie:

$a = 0,1234567 \cdot 10^0$

$b = 0,4711325 \cdot 10^4$

$c = -b$

avem:

$$fl(b+c) = 0,\ fl(a + fl(b+c)) = fl(a) = 0,1234567 \cdot 10^0$$

pe de o parte, iar pe de alta:

$$fl(a+b) = 0,00001234567 \cdot 10^4 + 0,4711325 \cdot 10^4 = 0,4711448 \cdot 10^4$$

$$fl(fl(a+b)+c) = (0,4711448 - 0,4711325))10^4 = 0,0000123 \cdot 10^4 = \\ = 0,1230000 \cdot 10^0.$$

§ 6. Propagarea erorilor

Dacă $x_1 = 2,32 \pm 0,03$ și $x_2 = 1,24 \pm 0,02$ atunci care este marginea erorii pentru $x_1 - x_2$?

Cea mai mare valoare posibilă a lui x_1 este 2,35 iar cea mai mică valoare posibilă a lui x_2 este 1,22, deci cea mai mare valoare posibilă a lui $x_1 - x_2$ este $2,35 - 1,22 = 1,13$.

Similar cea mai mică valoare a diferenței $x_1 - x_2$ este 1,03

Deci:

$$1,03 \leqslant x_1 - x_2 \leqslant 1,13$$

sau

$$x_1 - x_2 = 1,08 \pm 0,05.$$

În cazul general dacă $x_1 = \widetilde{x}_1 \pm \varepsilon_1$, $x_2 = \widetilde{x}_2 \pm \varepsilon_2$ avem:

$$\widetilde{x}_1 - \varepsilon_1 - (\widetilde{x}_2 + \varepsilon_2) \leqslant x_1 - x_2 \leqslant \widetilde{x}_1 + \varepsilon_1 - (\widetilde{x}_2 - \varepsilon_2),$$

$$\widetilde{x}_1 - \widetilde{x}_2 - (\varepsilon_1 + \varepsilon_2) \leqslant x_1 - x_2 \leqslant \widetilde{x}_1 - \widetilde{x}_2 + (\varepsilon_1 + \varepsilon_2),$$

$$x_1 - x_2 = \widetilde{x}_1 - \widetilde{x}_2 \pm (\varepsilon_1 + \varepsilon_2).$$

Prin inducție ajungem la următoarea:

Regulă. La adunare și scădere, marginile pentru eroarea absolută sînt date de suma marginilor erorilor absolute în operanzi.

Conform definiției erorii relative (vezi § 3) avem:

$$\frac{\widetilde{x} - x}{x} = r \text{ (unde } r \text{ este eroarea relativă)}$$

deci: $\widetilde{x} = x(1 + r)$.

Dacă $\widetilde{x}_1$ și $\widetilde{x}_2$ au erorile relative r_1 și r_2 avem:

$$\widetilde{x}_1\widetilde{x}_2 = x_1(1 + r_1)\ x_2(1 + r_2) = x_1x_2(1 + r_1)\ (1 + r_2).$$

Urmează că eroarea relativă în produsul $\widetilde{x}_1\widetilde{x}_2$ este dată de relația:

$$\frac{\widetilde{x}_1\widetilde{x}_2 - x_1x_2}{x_1x_2} = (1 + r_1)\ (1 + r_2) - 1 = r_1 + r_2 + r_1r_2.$$

Dacă $|r_1| \ll 1$ și $|r_2| \ll 1$ atunci putem neglija termenul r_1r_2 și obținem eroarea relativă în $\widetilde{x}_1\widetilde{x}_2$ ca fiind aproximativ $r_1 + r_2$.

De exemplu *dacă* $r_1 = 0,03$ și $r_2 = -0,02$, *eroarea relativă în produs este:*

$$r_1 + r_2 + r_1r_2 = 0,03 - 0,02 - 0,0006 = 0,0094 \approx 0,01.$$

Să procedăm la fel pentru a găsi eroarea relativă la împărțire.

Avem:

$$\frac{\widetilde{x}_1 - x_1}{x_1} = r_1 \text{ și } \frac{\widetilde{x}_2 - x_2}{x_2} = r_2$$

deci:

$$\widetilde{x}_1 = x_1(1 + r_1)$$

$$\widetilde{x}_2 = x_2(1 + r_2)$$

de unde:

$$\frac{\widetilde{x}_1}{\widetilde{x}_2} = \frac{x_1}{x_2} \cdot \frac{1 + r_1}{1 + r_2},$$

iar

$$\frac{\dfrac{\widetilde{x}_1}{\widetilde{x}_2} - \dfrac{x_1}{x_2}}{\dfrac{x_1}{x_2}} = \frac{1 + r_1}{1 + r_2} - 1 = \frac{r_1 - r_2}{1 + r_2}.$$

Dacă $|r_1| \ll 1$ și $|r_2| \ll 1$ obținem:

$$\frac{r_1 - r_2}{1 + r_2} \approx r_1 - r_2.$$

Dacă marginile erorilor relative în x_1, x_2, sînt ρ_1 și ρ_2 atunci $\rho_1 + \rho_2$ este cea mai bună margine superioară pentru $|r_1 + r_2|$ ca și pentru $|r_1 - r_2|$.
În concluzie avem următoarea regulă:
La înmulțire și la împărțire marginile pentru erorile relative în operanzi se adună.

O situație des întîlnită în practică, care conduce la rezultate eronate este următoarea:

Să presupunem că avem de făcut o scădere între doi operanzi, în care diferența este mult mai mică decît fiecare dintre ei.

Fie $y = x_1 - x_2$. Să notăm erorile în x_1 și x_2 cu Δx_1 și Δx_2 atunci:

$$|\Delta y| \leqslant |\Delta x_1| + |\Delta x_2|$$

deci

$$\frac{|\Delta y|}{y} \leqslant \frac{|\Delta x_1| + |\Delta x_2|}{|x_1 - x_2|}.$$

Aplicație: *Dacă* $x_1 = 0{,}6724 \pm \frac{1}{2}\ 10^{-4}$ *și*

$$x_2 = 0{,}6723 \pm \frac{1}{2}\ 10^{-4}$$

atunci:

$$x_1 - x_2 = 0{,}0001 \pm 0{,}0001.$$

Această anulare a rezultatului poate fi ocolită printr-o scriere convenabilă a formulelor de calcul sau aplicarea unui alt algoritm.

Să considerăm următorul exemplu:
ecuația $x^2 - 56x + 1 = 0$ *are rădăcinile:*

$$x_1 = 28 - \sqrt{783} \approx 28 - 27{,}982 = 0{,}018 \pm \frac{1}{2}\ 10^{-3}.$$

$$x_2 = 28 + \sqrt{783} \approx 55{,}982 \pm \frac{1}{2}\ 10^{-3}.$$

Se observă că deși rădăcina pătrată s-a calculat cu cinci cifre semnificative, totuși în prima rădăcină (x_1) primim numai două cifre semnificative iar în x_2 eroarea relativă este mai mică decît 10^{-5}.

Cum produsul rădăcinilor este $x_1 x_2 = 1$, putem utiliza următoarea cale pentru calculul lui x_1:

$$x_1 = \frac{1}{x_2} = \frac{1}{55{,}982} = 0{,}01786288$$

cu o eroare relativă mai mică decît 10^{-5}.

Așadar $x_1 = 0{,}0178629 \pm 0{,}0000002$ și astfel cu aceeași estimare a rădăcinii pătrate obținem pentru x_1 cinci cifre semnificative (în loc de două) utilizînd informația dată de produsul rădăcinilor unei ecuații de gradul al doilea.

În general dacă $|\delta| \ll x$ atunci vom scrie:

$$\sqrt{x+\delta} - \sqrt{x} = \frac{x+\delta-x}{\sqrt{x+\delta}+\sqrt{x}} = \frac{\delta}{\sqrt{x+\delta}+\sqrt{x}}.$$

De asemenea în multe situații este preferabil să utilizăm pentru calculul diferenței:

$$\cos(x+\delta) - \cos x$$

produsul,

$$-2 \sin\frac{\delta}{2} \sin\left(x + \frac{\delta}{2}\right).$$

Dăm acum cîteva reguli simple în vederea executării unui calcul practic mai precis.

I) **Cînd se adună sau se scad numere** (un șir de numere) **să se înceapă cu cele mai mici.**

II) **Să se evite pe cît posibil cu putință scăderea a două numere pozitive aproximativ egale. Deseori, după cum am văzut, putem rescrie formula de calcul pentru a evita acest lucru.**

III) **O expresie de forma** $a(b-c)$ **poate fi scrisă** $ab - ac$ **iar** $(a-b)/c$ **ca** $a/c - b/c$. **Dacă numerele sînt aproximativ egale să se facă scăderea înainte de înmulțire evitînd astfel erori suplimentare de rotunjire.**

IV) **Pe cît posibil să se minimizeze numărul operațiilor aritmetice.**

Exerciții

1. Scrieți în baza 10 numerele $(1000)_2$, $(100000)_8$, $(A64)_{16}$.

2. Fie numărul 1/3 de memorat într-o celulă de memorie cu 4 poziții zecimale. Cum apare acest număr în calculator? Cu ce eroare? Care este corecția? Același lucru pentru numerele π și $\sqrt{2}$.

3. În cazul reprezentării în calculator a numărului 1/3 (indiferent pe cîte poziții zecimale (în număr finit)), coincide eroarea de rotunjire cu eroarea de trunchiere?

4. Pentru reprezentarea (cu rotunjire prin tăiere) a unui număr irațional în calculator, coincide eroarea de rotunjire cu cea de trunchiere?

5. Pentru un calculator ce lucrează cu 7 poziții zecimale, care este unitatea?

6. Un calculator are rezervate 2 poziții zecimale pentru exponent. De ce se adaugă, în reprezentarea în calculator, numărul 50 exponentului real? Cum vor fi reprezentați pe aceste două poziții exponenții $-7, 6, -22, -11, 20$?

7. Avem un calculator cu 8 poziții binare pentru exponent. Care este cel mai mare număr întreg ce „încape“ în cele 8 poziții binare? Care este jumătatea acestui număr? Care este numărul ce se va adăuga exponentului real în reprezentarea în virgulă mobilă în calculator? În ce interval se situează exponentul real?

8. Fie de efectuat următoarele operații:

a) $2,1350 \cdot \pi$; *b*) $0,2567 \cdot e$; *c*) $\pi \cdot e$; d) $0,2135-0,212$; *e*) $1,2137+22,2137$ pe un calculator cu două poziții zecimale pentru exponent și 4 poziții zecimale pentru mantisă.

— Scrieți numerele în reprezentarea virgulă mobilă (normalizat) cu rotunjire.

— Efectuați operațiile pe celula simplă de memorie.

— Renormalizați.

— Efectuați calculele pe celulă dublă de memorie.

— Renormalizați.

9. Determinați eroarea relativă maximă pentru x, unde

$$x_1 = 2,3 \pm 0,2,$$

$$x_2 = 22,3 \pm 0,2,$$

$$x_3 = 203,3 \pm 0,2.$$

Care din aceste trei numere are precizia relativă cea mai bună?

10. Calculați eroarea relativă maximă pentru:

$$E = \frac{x_1 \cdot x_2^2}{x_3},$$

unde x_1, x_2, x_3 sînt cei de la punctul precedent.

11. Să se efectueze, cu precizia cea mai bună, adunarea:

$$S = x_1 + x_2 + x_3,$$

în virgulă mobilă cu 5 zecimale (exponent 2 poziții zecimale) pentru

$$x_1 = 0,11283,$$

$$x_2 = 0,000043888,$$

$$x_3 = 0,0082798.$$

12. Calculați:

$$y_n = \int_0^1 \frac{x^n}{x+5}\, dx,$$

pentru $n = 0, 1, 2, ..., 8$, utilizînd formula de recurență:

$$y_n + 5y_{n-1} = \frac{1}{n}.$$

a) Cum rezultă formula de recurență?

b) Utilizați trei zecimale în calcul și rotunjiți.

c) Arătați că $y_3 > y_2$ și $y_4 < 0$ ceea ce este evident absurd. Explicați de ce?

13. Utilizați în exemplul precedent formula:

$$y_{n-1} = \frac{1}{5n} - \frac{y_n}{5}.$$

Considerînd că $y_{10} \approx y_9$ arătați că $y_8 \approx 0{,}019$ și calculați $y_7, y_6, \ldots, y_0$.

Capitolul II

REZOLVAREA SISTEMELOR DE ECUAȚII ALGEBRICE LINIARE

§ 1. Punerea problemei. Condiționarea sistemelor

Să considerăm sistemul de două ecuații liniare algebrice cu două necunoscute:

$$\begin{cases} a_{11}x + a_{12}y = b_1 \\ a_{21}x + a_{22}y = b_2. \end{cases} \tag{1}$$

Coeficienții necunoscutelor și termenii liberi îi considerăm numere reale. Vom nota cu A matricea coeficienților necunoscutelor, deci:

$$A = \begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix}$$

și să presupunem că determinantul matricei A (îl notăm cu $\det A$) este diferit de zero, deci sistemul (1) admite soluție unică.

În această situație matricea A admite o inversă pe care o vom nota cu A^{-1}.

Avem:

$$A^{-1} = \begin{Vmatrix} \dfrac{a_{22}}{\det A} & \dfrac{-a_{12}}{\det A} \\ \\ \dfrac{-a_{21}}{\det A} & \dfrac{a_{11}}{\det A} \end{Vmatrix}$$

iar,

$$A \cdot A^{-1} = \begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix} \cdot \begin{Vmatrix} \dfrac{a_{22}}{\det A} & \dfrac{-a_{12}}{\det A} \\ \\ \dfrac{-a_{21}}{\det A} & \dfrac{a_{11}}{\det A} \end{Vmatrix} =$$

$$= \begin{Vmatrix} \dfrac{a_{11}a_{22} - a_{12}a_{21}}{\det A} & 0 \\ \\ 0 & \dfrac{-a_{21}a_{12} + a_{22}a_{11}}{\det A} \end{Vmatrix} = \begin{Vmatrix} 1 & 0 \\ 0 & 1 \end{Vmatrix}$$

Deci $A \cdot A^{-1} = I$ unde I este matricea unitate (la fel se verifică $A^{-1} \cdot A = I$).

Definiție. — Se numește lungimea euclidiană a matricei A (o vom nota cu $\|A\|_E$), **numărul pozitiv**

$$\|A\|_E = (|a_{11}|^2 + |a_{12}|^2 + |a_{21}|^2 + |a_{22}|^2)^{1/2}.$$

Calculînd $\|A^{-1}\|_E$ obținem:

$$\|A^{-1}\|_E = \frac{(|a_{11}|^2 + |a_{12}|^2 + |a_{21}|^2 + |a_{22}|^2)^{1/2}}{|\det A|}.$$

Deci:

$$\|A\|_E \cdot \|A^{-1}\|_E = \frac{|a_{11}|^2 + |a_{12}|^2 + |a_{21}|^2 + |a_{22}|^2}{|\det A|}. \tag{2}$$

Definiție. — Numărul pozitiv dat de relația (2) **se numește numărul de condiționare euclidiană a sistemului** (1) (îl vom nota $\text{cond}_E(A) = \|A\|_E \cdot \|A^{-1}\|_E$).

> *Observație.* Există și alte numere de condiționare pentru sisteme liniare. Despre unele vom vorbi în acest context iar despre altele vor afla cei ce vor intra în studiul aprofundat al tehnicii de calcul.

Din punct de vedere teoretic sistemul (1) admite soluție unică dacă $\det A \neq 0$. Din punct de vedere al calculului practic lucrurile nu stau deloc astfel.

Să considerăm următorul sistem de ecuații, pe care ne propunem să-l rezolvăm considerînd coeficienții săi dați cu două cifre deci,

$$\begin{cases} 8x + 9y = 17 \\ 9x + 10y = 19 \end{cases} \tag{3}$$

Cu regula lui *Cramer* rezultă imediat că unica soluție a sistemului este: $x = 1$, $y = 1$.

Din punct de vedere geometric, a rezolva sistemul (3) revine la a intersecta două drepte (fig. II.1).

Considerînd $x = 2{,}121$, $y = 0$, în (3) obținem:

$$\begin{cases} 8x + 9y = 16{,}968 \\ 9x + 10y = 19{,}089. \end{cases} \tag{4}$$

Prin rotunjire, la două cifre semnificative, membrul drept al sistemului (4) corespunde cu membrul drept al sistemului (3). În concluzie, soluția $x = 2{,}121$, $y = 0$ trebuie acceptată ca fiind la fel de bună ca cea unică întrucît coeficienții și termenii liberi ai sistemului inițial au fost dați cu două cifre.

Din figura dată se vede că dreptele sînt aproape confundate, iar punctul $x = 2{,}121$, $y = 0$ nu se găsește pe nici una din cele două drepte, dar el se găsește foarte aproape de amîndouă.

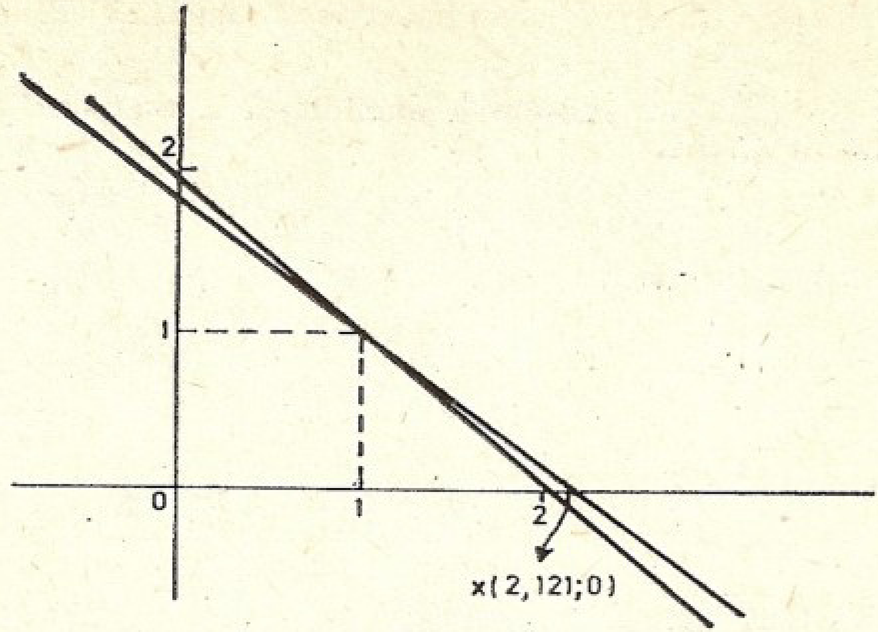


Fig. II. 1

Calculînd numărul de condiționare al matricei asociate sistemului (3) obținem:

$$\mathrm{cond}_E(A) = \frac{64 + 81 + 81 + 100}{1} = 326.$$

După cum observăm acest număr este destul de „mare" în sensul următor: dacă efectuăm calculul în rezolvarea sistemului (4) cu trei zecimale exacte, atunci obținem soluția $x = 2{,}121$, $y = 0$, pe cînd rotunjind în (4) la două cifre avem soluția: $x = 1$, $y = 1$.

Așadar numărul de condiționare ne dă o oarecare imagine despre sensibilitatea soluției unui sistem liniar algebric la perturbații mici în membrul drept sau eventual în coeficienți.

În fapt, problema rezolvării sistemului (4) poate fi privită și în felul următor: la o perturbare foarte mică a membrului drept, vectorul

$$\begin{Vmatrix} 17 \\ 19 \end{Vmatrix} \text{ devenind } \begin{Vmatrix} 16{,}968 \\ 19{,}098 \end{Vmatrix}$$

în soluție avem o perturbare destul de mare căci vectorul

$$\begin{Vmatrix} 1 \\ 1 \end{Vmatrix} \text{ a devenit } \begin{Vmatrix} 2{,}121 \\ 0 \end{Vmatrix}.$$

Să introducem acum pentru sistemul (1) și un alt număr de condiționare. Anume asociem matricei A, numărul pozitiv:

$$\|A\|_\infty = \max\left(|a_{11}| + |a_{12}|;\ |a_{21}| + |a_{22}|\right)$$

(cu alte cuvinte valoarea maximă dintre numerele pozitive obținute prin adunarea pe linii a modulelor elementelor matricei A).

Observație. — Se remarcă faptul că și acest număr pozitiv definește o lungime pentru matricea A.

De asemenea asociind și matricei inverse, după aceeași regulă, numărul pozitiv $\|A^{-1}\|_\infty$, numărul de condiționare a matricei A (pe care-l vom nota cu $\text{cond}_\infty(A)$) este prin definiție produsul numerelor pozitive $\|A\|_\infty$ și $\|A^{-1}\|_\infty$ deci:

$$\|A\|_\infty \cdot \|A^{-1}\|_\infty = \text{cond}_\infty (A).$$

De exemplu pentru sistemul (3) *avem:*

$$\|A\|_\infty = \max\ (17;\ 19) = 19.$$

Cum $A^{-1} = \begin{Vmatrix} -10 & 9 \\ 9 & -8 \end{Vmatrix}$ avem:

$$\|A^{-1}\|_\infty = \max\ (|-10| + |9|;\ |9| + |-8|) = 19.$$

Rezultă că:

$$\text{cond}_\infty (A) = 19 \cdot 19 = 361.$$

Observăm că și acest număr de condiționare este de același ordin de mărime ca și $\text{cond}_E (A)$.

Metodele de rezolvare a sistemelor liniare algebrice pe care le dezvoltăm aici se justifică prin faptul că metoda lui Cramer de exemplu este foarte neeconomicoasă (calculul unui determinant necesită foarte multe operații) și practic utilizabilă numai pentru sisteme de două sau trei ecuații cu tot atîtea necunoscute.

Problemele practice dau naștere la sisteme de ecuații de ordinul sutelor sau chiar miilor de necunoscute și atunci au trebuit căutate metode eficiente.

Vom descrie în cele ce urmează, așa-zisele **metode directe** (datorate lui *Gauss*) adică metode prin care soluția se obține într-un număr finit de pași (un număr finit de operații).

Apoi ne vom ocupa, pe scurt, de **metodele iterative** (indirecte) în care soluția sistemului se obține printr-un **proces de trecere la limită.**

Evident în metodele iterative nu putem efectua o infinitate de operații, deci ne oprim după calculul unui număr finit de aproximante (se face în acest caz o eroare de trunchiere).

În metodele directe, dacă s-ar lucra ideal (deci fără erori de rotunjire), lucru imposibil dealtfel, s-ar obține soluția exactă.

În timp ce în metodele directe erorile de rotunjire devin foarte periculoase, conducînd la soluții fără sens dacă sistemul este rău condiționat, în metodele iterative, în ciuda erorii de trunchiere erorile de rotunjire nu se acumulează.

§ 2. Metoda eliminării a lui Gauss

Pentru ilustrarea acestei metode directe vom utiliza un sistem de trei ecuații liniare algebrice cu trei necunoscute, despre care presupunem că admite soluție unică:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3. \end{cases} \tag{5}$$

Cel puțin unul din coeficienții a_{11}, a_{21}, a_{31} este diferit de zero, căci altfel sistemul (5) n-ar mai fi cu trei necunoscute. Dacă a_{11} este egal cu 0 rearanjăm ecuațiile astfel încît coeficientul lui x_1 din prima ecuație să nu fie zero.

Soluția sistemului rămîne, evident, neschimbată dacă schimbăm ecuațiile între ele (De ce?).

Definim un **multiplicator** $m_2 = \dfrac{a_{21}}{a_{11}}$.

Înmulțim prima ecuație din sistem cu m_2 și o scădem din ecuația a doua (atenție acum „prima" și a „doua" se referă la ecuațiile deja rearanjate dacă a fost cazul).

După efectuarea operației de scădere, ecuația a doua din sistem va fi:

$$(a_{21} - m_2a_{11})x_1 + (a_{22} - m_2a_{12})x_2 + (a_{23} - m_2a_{13})x_3 =$$
$$= b_2 - m_2b_1. \tag{6}$$

Dar:

$$a_{21} - m_2a_{11} = a_{21} - \frac{a_{21}}{a_{11}}a_{11} = 0$$

deci x_1 a fost eliminat din ecuația a doua $\left(\text{acum se observă de ce a fost ales } m_2 = \dfrac{a_{21}}{a_{11}}\right)$.

Notăm:

$$a'_{22} = a_{22} - m_2a_{12},$$
$$a'_{23} = a_{23} - m_2a_{13},$$
$$b'_2 = b_2 - m_2b_1.$$

și astfel (6) devine:

$$a'_{22}x_2 + a'_{23}x_3 = b'_2. \tag{7}$$

Deci a doua ecuație din sistemul (5) va fi dată de relația (7).

Soluția acestui nou sistem este aceeași cu soluția sistemului inițial.

Similar, definim $m_3 = \dfrac{a_{31}}{a_{11}}$ și înmulțim prima ecuație cu m_3 și o scădem din a treia. Eliminăm din nou coeficientul lui x_1 și obținem:

$$a'_{32}x_2 + a'_{33}x_3 = b'_3 \tag{8}$$

unde:

$$a'_{32} = a_{32} - m_3a_{12},$$
$$a'_{33} = a_{33} - m_3a_{13},$$
$$b'_3 = b_3 - m_3b_1.$$

Înlocuind în sistemul inițial a doua ecuație prin (7) și a treia ecuație prin (8) obținem sistemul:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ \qquad\quad a'_{22}x_2 + a'_{23}x_3 = b'_2 \\ \qquad\quad a'_{32}x_2 + a'_{33}x_3 = b'_3. \end{cases} \tag{9}$$

Sistemul obținut are aceeași soluție ca sistemul inițial și prezintă avantajul că x_1 nu mai apare în ultimele două ecuații.

Să trecem acum la eliminarea lui x_2 din ultimele două ecuații.

Dacă $a'_{22} = 0$, permutăm ultimele două ecuații. Dacă și a'_{22} și a'_{32} sînt egali cu zero, am avea o infinitate de soluții sau nici una, ceea ce nu este posibil deoarece am admis că sistemul inițial are soluție unică.

Alegem multiplicatorul $m'_3 = \dfrac{a'_{32}}{a'_{22}}$.

Dacă înmulțim a doua ecuație din sistemul (9) cu m'_3 și o scădem din a treia, obținem:

$$(a'_{32} - m'_3 \cdot a'_{22}) \cdot x_2 + (a'_{33} - m'_3 \cdot a'_{23}) \cdot x_3 = b'_3 - m'_3 \cdot b'_2.$$

Din nou $a'_{32} - m'_3 \cdot a'_{22} = 0$, și notînd

$$a''_{33} = a'_{33} - m'_3 \cdot a'_{23}$$

$$b''_3 = b'_3 - m'_3 \cdot b'_2$$

obținem:

$$a''_{33} \cdot x_3 = b''_3. \tag{10}$$

Astfel putem înlocui sistemul (9) prin sistemul echivalent:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ \qquad\quad a'_{22}x_2 + a'_{23}x_3 = b'_2 \\ \qquad\qquad\qquad a''_{33}x_3 = b''_3. \end{cases} \tag{11}$$

Matricea asociată coeficienților necunoscutelor sistemului (11) este superior triunghiulară:

$$U = \begin{Vmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a'_{22} & a'_{23} \\ 0 & 0 & a''_{33} \end{Vmatrix}$$

și de aceea sistemul (11) se numește **superior triunghiular.**

Sistemul (11) se rezolvă foarte simplu, prin substituție, începînd cu ecuația a treia.

Deci:

$$x_3 = \frac{b''_3}{a''_{33}}, \quad x_2 = \frac{b'_2 - a'_{23}x_3}{a'_{22}},$$

$$x_1 = \frac{b_1 - a_{12}x_2 - a_{13}x_3}{a_{11}}.$$

Amintim că $a_{11} \neq 0$, $a'_{22} \neq 0$ iar a''_{33} este de asemenea diferit de zero deoarece în caz contrar ar rezulta că matricea asociată coeficienților sistemului (5) ar fi singulară (cu determinantul zero).

În ecuația a doua și a treia din (11) renotăm coeficienții cu a_{22}, a_{23}, b_2 respectiv a_{33}, b_3. Aceste ecuații se scriu:

$$a_{22}x_2 + a_{23}x_3 = b_2$$
$$a_{33}x_3 = b_3.$$

Deși prin această renotare nu mai putem identifica pe a_{22}, a_{23}, b_2 respectiv a_{33}, b_3 cu coeficienții ecuației doi și trei din (5) renotarea este posibilă deoarece nu mai folosim niciodată acest sistem. Renotarea prezintă avantaj la scrierea programului de calcul pentru a nu introduce noi variabile ci folosirea unor locații din program rămase libere.

Operațiile folosite în rezolvarea sistemului (5) asupra coeficienților și termenului liber nu modifică soluția sistemului.

Acum să trecem la descrierea generală a **algoritmului lui Gauss.**

Să considerăm sistemul liniar de n ecuații cu n necunoscute, de asemenea presupunem că sistemul admite soluție unică, deci că matricea coeficienților necunoscutelor este nesingulară:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2j}x_j + \ldots + a_{2n}x_n = b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n = b_i \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nj}x_j + \ldots + a_{nn}x_n = b_n. \end{cases} \tag{12}$$

Dacă notăm:

$$A = \begin{Vmatrix} a_{11} & a_{12} \ldots a_{1j} \ldots a_{1n} \\ a_{21} & a_{22} \ldots a_{2j} \ldots a_{2n} \\ \ldots & \ldots\ldots\ldots\ldots \\ a_{n1} & a_{n2} \ldots a_{nj} \ldots a_{nn} \end{Vmatrix}$$

$$x = \begin{Vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Vmatrix}; \quad b = \begin{Vmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{Vmatrix}$$

atunci sistemul (12) se poate scrie în forma matriceală:

$$Ax = b. \tag{13}$$

Presupunem că $a_{11} \neq 0$. Eliminăm pe x_1 din ultimele $(n-1)$ ecuații scăzînd din ecuația „i" prima ecuație multiplicată cu:

$$m_{i1} = \frac{a_{i1}}{a_{11}}, \quad i = 2, 3, 4, \ldots, n$$

deci avem $n-1$ multiplicatori).

Ultimele $n-1$ ecuații nu vor mai conține necunoscuta x_1; vor deveni:

$$\begin{cases} a_{22}^{(2)}x_2 + \dots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ a_{n2}^{(2)}x_2 + \dots + a_{nn}^{(2)}x_n = b_n^{(2)} \end{cases}$$

unde:

$$a_{ij}^{(2)} = a_{ij} - m_{i1}a_{1j}$$

$$b_i^{(2)} = b_i - m_{i1}b_1,$$

$$i = 2,\ 3,\ \dots,\ n.$$

Acest sistem este un sistem de $n-1$ ecuații în necunoscutele x_2, x_3, $x_4, \dots$ $\dots, x_n$.

Dacă $a_{22}^{(2)} \neq 0$ similar putem elimina pe x_2 din ultimele $(n-2)$ ecuații în necunoscutele $x_2, x_3, \dots, x_n$.

Fie:

$$m_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}},\quad i = 3,\ 4,\ \dots,\ n$$

coeficienții acestui sistem sînt dați de:

$$a_{ij}^{(3)} = a_{ij}^{(2)} - m_{i2}a_{2j}^{(2)}$$

$$b_i^{(3)} = b_i^{(2)} - m_{i2}b_2^{(2)}, \qquad i = 3,\ 4,\ \dots,\ n.$$

Elementele a_{11}, $a_{22}^{(2)}$, $a_{33}^{(3)}$, ... ce apar în timpul procesului de eliminare a necunoscutelor poartă numele de **elemente pivot.**

Dacă toate elementele pivot sînt diferite de zero, după $n-1$ pași obținem o singură ecuație:

$$a_{nn}^{(n)}x_n = b_n^{(n)}.$$

Acum, colectînd prima ecuație din fiecare pas obținem:

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \dots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ \qquad\qquad a_{22}^{(2)}x_2 + \dots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \qquad\qquad\qquad\qquad\qquad a_{nn}^{(n)}x_n = b_n^{(n)} \end{cases} \tag{14}$$

unde am notat: $a_{ij}^{(1)} = a_{ij}$; $b_i^{(1)} = b_i$ pentru coeficienții sistemului inițial pe care dorim să-l rezolvăm prin metoda Gauss.

Am folosit în (14) o notație mai complicată pentru a arăta exact cum se transformă coeficienții și termenul liber în metoda lui Gauss. Remarcăm din nou, ca și în cazul sistemelor de trei ecuații, că la scrierea programului de calcul este mai simplu să revenim la vechea scriere a ecuației.

Sistemul (14) este un sistem triunghiular de forma:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1,n-1}x_{n-1} + a_{1n}x_n = b_1 \\ \qquad a_{22}x_2 + \ldots + a_{2,n-1}x_{n-1} + a_{2n}x_n = b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ \qquad\qquad\qquad a_{nn}x_n = b_n. \end{cases} \tag{15}$$

Dacă presupunem că $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$ atunci necunoscutele x_n, $x_{n-1}, \ldots, x_1$ se determină astfel:

$$\begin{aligned} x_n &= \frac{b_n}{a_{nn}} \text{ (din ultima ecuație)} \\ x_{n-1} &= \frac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}} \text{ (din penultima ecuație)} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_1 &= \frac{b_1 - a_{1n}x_n - a_{1,n-1}x_{n-1} - \ldots - a_{12}x_2}{a_{11}}. \end{aligned} \tag{16}$$

Relațiile (16) se pot scrie compact astfel:

$$x_i = \frac{b_i - \sum_{k=i+1}^{n} a_{ik}x_k}{a_{ii}}, \quad i = n, n-1, \ldots, 1. \tag{17}$$

Matricea coeficienților sistemului (15) *se numește superior triunghiulară.* În cazul în care am avea o *matrice inferior triunghiulară, deci cu zerouri deasupra diagonalei principale,* rezolvarea acestui sistem este aceeași numai că se vor scoate necunoscutele în ordinea $x_1, x_2, \ldots, x_n$.

Dacă studiem formula (17) observăm că pentru rezolvarea unui sistem triunghiular se fac „n“ împărțiri și

$$\sum_{i=1}^{n} (i-1) = \frac{1}{2} \cdot n(n-1) \text{ adunări și înmulțiri.}$$

Dacă „n“ este „mare“ atunci $n^2 \gg n$ și

$$\frac{1}{2} \cdot n(n-1) \approx \frac{1}{2} n^2.$$

Evaluăm în calcul numărul de operații pentru a ne da seama cît timp calculator necesită o anume problemă și deci cît costă rezolvarea ei.

§ 3. Formularea matriceală a algoritmului lui Gauss

Urmărind cu atenție algoritmul lui Gauss se observă că operațiile care se fac asupra coeficienților a_{ij} se fac și asupra componentelor vectorului format cu termenii liberi.

Formăm matricea $A^{\&}$ (extinsa matricii A) din matricea A adăugîndu-i o coloană formată din vectorul b.

Deci:

$$A^{\&}=\begin{Vmatrix} a_{11} & a_{12} \dots a_{1j} \dots a_{1n} & b_1 \\ \dots & \dots & \dots \\ a_{i1} & a_{i2} \dots a_{ij} \dots a_{in} & b_i \\ \dots & \dots & \dots \\ a_{n1} & a_{n2} \dots a_{nj} \dots a_{nn} & b_n \end{Vmatrix} \tag{18}$$

Operațiile din algoritmul Gauss ce se efectuează cu liniile din matricea $A^{\&}$ sînt:

— înmulțirea unei linii cu o constantă,

— scăderea unei linii din alta și înlocuirea celei de a doua cu rezultatul scăderii,

— rearanjarea liniilor.

Dacă notăm cu $a_{i,n+1}$ pe b_i atunci matricea $A^{\&}$ cu „n" linii și „$n+1$" coloane este:

$$A^{\&}=\begin{Vmatrix} a_{11} & a_{12} \dots a_{1n} & a_{1,n+1} \\ a_{21} & a_{22} \dots a_{2n} & a_{2,n+1} \\ \dots & \dots & \dots \\ a_{n1} & a_{n2} \dots a_{nn} & a_{n,n+1} \end{Vmatrix} \tag{19}$$

În urma operațiilor descrise în paragraful precedent matricea $A^{\prime}$ devine:

$$A^{\&}=\begin{Vmatrix} a_{11} & a_{12} \dots a_{1,n-1} & a_{1n} & a_{1,n+1} \\ 0 & a_{22} \dots a_{2,n-1} & a_{2n} & a_{2,n+1} \\ \dots & \dots & \dots & \dots \\ 0 & 0 \dots a_{n-1,n-1} & a_{n-1,n} & a_{n-1,n+1} \\ 0 & 0 \dots 0 & a_{nn} & a_{n,n+1} \end{Vmatrix} \tag{20}$$

Se observă procedeul de renotare a elementelor matricii $A^{\&}$ (vezi organigrama și programul) căci elementele a_{ij} din (20) se calculează după cum s-a arătat în paragraful precedent.

În procesul de eliminare se observă că la eliminarea necunoscutei x_k elementele matricei din coloana k și liniile $k+1, \dots, n (k=1, 2, \dots, n-1)$ devin nule.

Înmulțim acum ultima linie din matricea (20) cu $\frac{1}{a_{nn}}$ și obținem:

$$\left(0, 0, 0, \dots, 0, 1 \frac{a_{n,n+1}}{a_{nn}}\right). \tag{21}$$

Renotînd ultimul element al vectorului (21) cu $a_{n,n+1}$ obținem:

$$(0, 0, 0, \ldots, 0, 1, a_{n,n+1}). \tag{22}$$

Ultima componentă a vectorului (22) este componenta x_n a vectorului soluție.

Înmulțim pe (22) cu $a_{n-1,n}$ și scădem din linia „$n-1$" și astfel obținem:

$$(0, 0, 0, \ldots, a_{n-1,n-1}, 0, a_{n-1,n+1} - a_{n-1,n}a_{n,n+1}). \tag{23}$$

Înmulțind vectorul (23) cu $\frac{1}{a_{n-1,n-1}}$ obținem pe ultima poziție (deci în coloana $n+1$ a matricii $A^{\&}$) valoarea necunoscutei x_{n-1}.

Continuînd procedeul obținem pe rînd necunoscutele $x_{n-2}, x_{n-3}, \ldots, x_2, x_1$ și matricea $A^{\&}$ devine:

$$A^{\&} = \left\| \begin{matrix} 1 & & \mathbf{0} & a_{1,n+1} \\ & \ddots & & a_{2,n+1} \\ & & & \vdots \\ \mathbf{0} & & 1 & a_{n,n+1} \end{matrix} \right\| \tag{24}$$

Observație. — În matricea (24) elementele de pe ultima coloană sînt exprimate de relațiile (17) cu b_j înlocuit cu $a_{j,n+1}$.

§ 4. Organigrama pentru metoda lui Gauss și programul de calcul în FORTRAN

În paragrafele 2 și 3 am studiat metoda eliminării a lui Gauss și s-a arătat că pentru fiecare pas „i" în procesul de eliminare este necesar să avem $a_{ii} \neq 0$. Dacă $a_{ii} = 0$ se caută în coloana i un element diferit de zero, fie acesta a_{ik} și se schimbă linia k cu linia i. Dacă se folosește renotarea înseamnă că avem $a_{ii} \neq 0$.

În cursul acestui capitol (§ 6) se va arăta că pentru a obține o soluție a sistemului algebric liniar, cît mai apropiată de soluția exactă, nu e necesar numai ca $a_{ii} \neq 0$ ci trebuie să avem:

$$|a_{ii}| = \max_{i \leqslant k \leqslant n} |a_{ki}|$$

(elementul pivot să fie cel mai mare, în modul, în coloana i).

Algoritmul lui Gauss de rezolvare a sistemelor liniare algebrice descris în paragraful 2 poate fi urmărit cu ușurință în organigrama din figura II.2.

Urmează programul de calcul corespunzător, în limbaj FORTRAN în care sînt demarcate prin cartele de comentariu blocurile de calcul din figura II. 2.

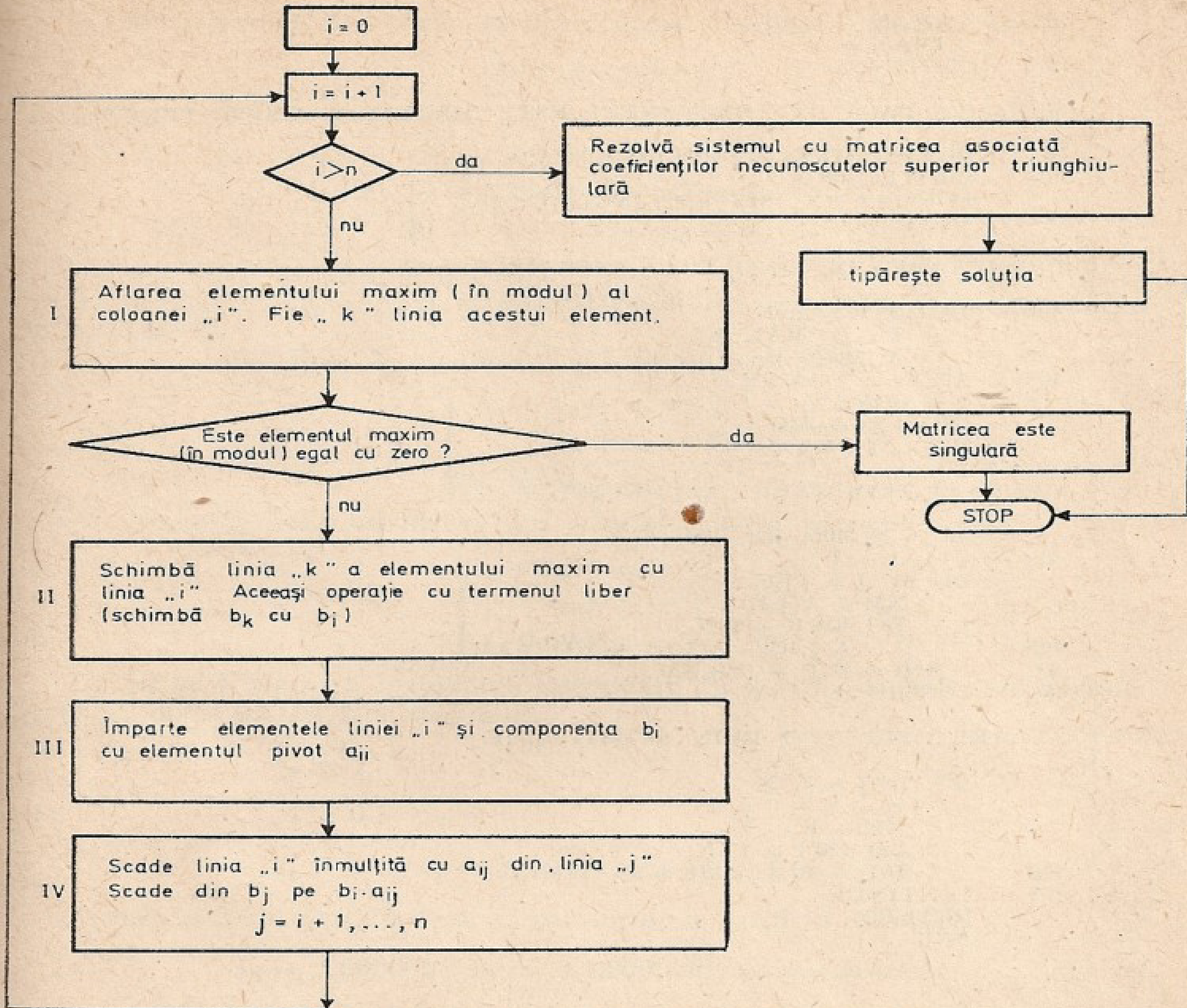


Fig. II. 2. Organigrama pentru metoda eliminării a lui Gauss

PROGRAMUL FORTRAN

pentru metoda eliminării a lui Gauss

```
      SUBROUTINE GAUSS (A,B,N,EPS)
      DIMENSION A(N,N), B(N)
C
C AICI ÎNCEPE PROCESUL DE ELIMINARE
C
          DO 100 I = 1, N
          AMAX = A(I,I)
          IMAX = I
C
C AFLAREA ELEMENTULUI MAXIM DIN COLOANA I
C
          DO 101 J = I, N
          IF(ABS(AMAX) — ABS(A(J,I))) 102, 101, 101
```

```
102       AMAX = A(J,I)
          IMAX = J
 101    CONTINUE
C
C VERIFICĂ DACĂ DETERMINANTUL ESTE DIFERIT DE ZERO
C
        IF(ABS(AMAX) - EPS) 103, 104, 104
 103    WRITE (108,3)
   3    FORMAT (1X, 'SISTEM SINGULAR')
        STOP
C
C PERMUTAREA DE LINII, DACĂ ESTE NECESARĂ
C
 104    DO 105 K = I, N
              T = A(IMAX, K)
              A(IMAX, K) = A(I,K)
 105    A(I,K) = T/AMAX
        T = B(IMAX)
        B(IMAX) = B(I)
        B(I) = T/AMAX
C
C URMEAZĂ ELIMINAREA NECUNOSCUTEI X(I)
C
        IF(I-N) 106, 100, 100
 106    JJ = I + 1
        DO 107 J = JJ,N
              XM = A(J,I)
              DO 108 K = I, N
 108                A(J,K) = A(J,K) - A(I,K)·XM
 107      B(J) = B(J) - B(I).XM
 100    CONTINUE
C
C AFLAREA SOLUŢIEI PRIN SUBSTITUŢIE
C
        DO 109 J = 2, N
              K = N - J + 1
              KK = K + 1
              DO 110 I = 1, K
 110          B(I) = B(I) - A(I, KK) · B(KK)
 109    CONTINUE
        RETURN
        END
```

Exemplu numeric
Fie sistemul:

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 + 1{,}0001\ \ x_2 + 2x_3 = 2 \\ x_1 + 2x_2 + 2x_3 = 1 \end{cases}$$

pe care dorim să-l rezolvăm prin metoda Gauss.

La pasul **1** în procesul de eliminare avem:

— elementul maxim în modul în coloana 1 este 1, deci elementul pivot va fi $a_{11} = 1$,

— întrucît elementul pivot este egal cu unitatea, prima ecuaţie a sistemului este deja normată (împărţită la elementul pivot),

— eliminarea necunoscutei x_1 din ecuaţia a doua (se scade prima ecuaţie din a doua, deoarece $a_{21} = 1$) ne conduce la:

$$0{,}0001\ x_2 + x_3 = 1,$$

— eliminarea necunoscutei x_1 din ecuația a treia ne conduce la:

$$x_2 + x_3 = 0.$$

Sistemul de ecuații devine:

$$\begin{cases} x_1 + \quad\quad\; x_2 + x_3 = 1 \\ 0 + 0{,}0001\, x_2 + x_3 = 1 \\ 0 + \quad\quad\; x_2 + x_3 = 0. \end{cases}$$

La pasul 2 al eliminării avem:

$a_{22} = 0{,}0001$; $a_{32} = 1$, și elementul pivot este a_{32} (este element maxim în coloana 2),

— se schimbă linia 2 cu linia 3 și sistemul devine:

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ \quad\quad x_2 + x_3 = 0 \\ 0{,}0001 x_2 + x_3 = 1 \end{cases}$$

— eliminăm necunoscuta x_2 din ecuația a treia și avem:

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ \quad\quad x_2 + x_3 = 0 \\ \quad 0{,}9999 x_3 = 1. \end{cases}$$

La pasul 3 urmează rezolvarea sistemului cu matrice superior triunghiulară (se operează cu trei zecimale)

$$x_3 = 1$$

$$x_2 = 0 - x_3 = -1$$

$$x_1 = 1 - x_2 - x_3 = 1 - (-1) - 1 = 1$$

Același sistem, rezolvat cu calculatorul, a furnizat soluția:

$$x_1 = 1{,}000000;\; x_2 = -1{,}000099;\; x_3 = 1{,}000099$$

(calculul s-a efectuat cu 7 zecimale).

§ 5. Strategia pivotării

În descrierea algoritmului lui Gauss noi am presupus (vezi §2) că elementele $a_{kk}^{(k)}$ (elementele pivot) sînt diferite de zero, altfel sistemul este singular.

Să considerăm următorul sistem de ecuații:

$$\begin{cases} x_1 + \; x_2 + \; x_3 = 1 \\ x_1 + \; x_2 + 2x_3 = 2 \\ x_1 + 2x_2 + 2x_3 = 1. \end{cases} \tag{25}$$

Sistemul admite soluție unică (determinantul matricei coeficienților este nenul) $x_1 = 1$, $x_2 = -1$, $x_3 = 1$.

Aplicînd procedeul lui Gauss, după primul pas primim:

$$\begin{aligned} x_3 &= 1 \\ x_2 + x_3 &= 0. \end{aligned} \tag{26}$$

Deci, observăm că $a_{22}^{(2)} = 0$ (coeficientul lui x_2 în prima ecuație (26) — vezi § 2).

Pentru a putea rezolva acest neajuns permutăm cele două ecuații și scriem sistemul (26) ca:

$$\begin{aligned} x_2 + x_3 &= 0 \\ x_3 &= 1 \end{aligned} \tag{27}$$

și acum obținem un sistem deja triunghiular care rezolvat ne conduce la soluția căutată.

Să considerăm acum cazul general. Deci la pasul k am obținut $a_{kk}^{(k)} = 0$. Atunci măcar unul din elementele $a_{k+1,k}^{(k)} a_{k+2,k}^{(k)} \ldots a_{nk}^{(k)}$ este nenul, altfel matricea A ar fi singulară.

Să notăm acel element cu $a_{rk}^{(k)} \neq 0$. Vom proceda atunci la schimbarea între ele a liniilor k și r apoi vom continua eliminarea.

Deci orice sistem nesingular poate fi redus la o formă triunghiulară cu ajutorul metodei lui Gauss combinată cu schimbări de linii între ele.

Să revenim acum la sistemul (25) în care coeficientul lui x_2 din ecuația a doua (a_{22}) îl înlocuim cu 1,0001. Rezolvînd din nou prin eliminare obținem:

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ 0{,}0001 x_2 + x_3 = 1 \\ -9\,999 x_3 = -10\,000. \end{cases} \tag{28}$$

Acum, rezolvînd sistemul (28) ca un sistem triunghiular și utilizînd trei zecimale obținem:

$$x_1 = 0, \quad x_2 = 0, \quad x_3 = 1{,}000. \tag{29}$$

Am scris 1,000 pentru a scoate în evidență că lucrăm cu trei zecimale.

Soluția exactă cu rotunjire la patru zecimale este:

$$x_1 = 1{,}000, \; x_2 = 1{,}0001, \quad x_3 = 1{,}0001.$$

Acest neajuns provine din faptul că $a_{22}^{(2)}$ a devenit după primul pas al eliminării 0,0001 (deci zero cînd lucrăm în trei zecimale), așadar un element pivot nu are voie să fie nici apropiat de zero.

Noi am rezolvat în § 4 sistemul:

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 + 1{,}0001 x_2 + 2 x_3 = 2 \\ x_1 + 2 x_2 + 2 x_3 = 1 \end{cases}$$

prin schimbarea liniilor 2 și 3 și am obținut soluția:

$$x_1 = -x_2 = x_3 = 1{,}000$$

corectă în trei zecimale (am utilizat trei zecimale în calculul soluției).

Exemplul prezentat ne arată că erorile de rotunjire pot avea efect catastrofal în metoda Gauss, de aceea foarte des la pasul k urmăm strategia (numită strategie de pivotare):

— căutăm cel mai mic număr „r" pentru care:

$$|a_{rk}^{(k)}| = \max |a_{ik}^{(i)}| \quad k \leqslant i \leqslant n$$

și schimbăm liniile k și r.

Cu alte cuvinte, pivotul la fiecare pas este cel mai mare element în modul în coloana considerată.

Deci, începem prin a căuta pivotul în prima coloană. Linia întîi și linia care conține elementul pivot (cel mai mare element în modul din coloana întîi) se schimbă între ele.

Procesul se repetă cu a doua coloană și așa mai departe.

Cînd linia ce conține pivotul este înmulțită cu un coeficient (înainte de a o scădea din altă linie) ea este de fapt înmulțită cu un număr subunitar (vezi definiția multiplicatorilor m_i § 2) și astfel erorile de rotunjire în liniile pivot sînt înmulțite cu un număr mai mic ca 1.

Observație. Procesul eliminării a lui Gauss are loc fără schimbări de linii între ele dacă matricea coeficienților necunoscutelor

$$A = \|a_{ij}\|_{\substack{i=\overline{1,n}\\ j=\overline{1,n}}}$$

are proprietatea că este **diagonal dominantă,** adică dacă:

$$|a_{ii}| \geqslant \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, 2, \ldots, n$$

deci elementele de pe diagonala principală sînt mai mari (sau egale) în modul decît suma modulelor elementelor din linia respectivă.

De exemplu:

$$\begin{Vmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{Vmatrix}$$

este o matrice *diagonal dominantă de ordinul trei.*

Exercițiu. *Este matricea de ordinul* 7×7 *diagonal dominantă?*

$$\begin{Vmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 \end{Vmatrix}$$

(Astfel de matrice se numesc tridiagonale și apar foarte frecvent în practică. Se numesc matrice rare, avînd multe elemente egale cu zero.)

§ 6. Sisteme rău condiționate și erorile de rotunjire

Dacă considerăm sistemul $Ax = b$ unde:

$$A = \begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix}; \qquad x = \begin{Vmatrix} x_1 \\ x_2 \end{Vmatrix}; \qquad b = \begin{Vmatrix} b_1 \\ b_2 \end{Vmatrix}$$

și $\bar{x}$ soluția sa calculată aproximativ, vom numi vector rezidual r, vectorul dat de: $r = b - A\bar{x}$.

Dacă $r = \begin{Vmatrix} 0 \\ 0 \end{Vmatrix}$ atunci $\bar{x}$ este soluția exactă, deci în mod natural ne-am aștepta că dacă cele două componente ale vectorului rezidual ar fi mici, atunci $\bar{x}$ să fie o soluție suficient de bună pentru sistemul considerat.

Lucrurile nu stau nici pe departe astfel:

Iată următorul exemplu foarte sugestiv:

Fie:

$$A = \begin{Vmatrix} 1,2969 & 0,8648 \\ 0,2161 & 0,1441 \end{Vmatrix},$$

$$b = \begin{Vmatrix} 0,8648 \\ 0,1440 \end{Vmatrix} \tag{31}$$

Presupunem că:

$$\bar{x} = \begin{Vmatrix} 0,9911 \\ -0,4870 \end{Vmatrix}$$

Atunci:

$$r = \begin{Vmatrix} r_1 \\ r_2 \end{Vmatrix} = \begin{Vmatrix} 0,8642 \\ 0,1440 \end{Vmatrix} - \begin{Vmatrix} 1,2969 & 0,8648 \\ 0,2161 & 0,1441 \end{Vmatrix} \cdot \begin{Vmatrix} 0,9911 \\ -0,4870 \end{Vmatrix} = \begin{Vmatrix} -10^{-8} \\ 10^{-8} \end{Vmatrix}.$$

Deci componentele lui r sînt destul de mici, de ordinul 10^{-8}. Totuși $\bar{x}$ cu componentele date este foarte departe de soluția exactă, care este:

$$x = \begin{Vmatrix} 2 \\ -2 \end{Vmatrix}.$$

Sistemul considerat este foarte rău condiționat. După eliminarea lui x_1 obținem:

$$a_{22}^{(2)} x_2 = b_2^{(2)}, \text{ unde:}$$

$$a_{22}^{(2)} = 0,1441 - \frac{0,2161}{1,2969} \cdot 0,8648 = 0,1441 - 0,1440999923 \approx 10^{-8}.$$

Deci o mică schimbare în coeficientul 0,1441 provoacă o mare schimbare în $a_{22}^{(2)}$, deci în x_2.

Dacă inversăm matricea A, obținem:

$$A^{-1} = 10^8 \begin{Vmatrix} 0,1441 & -0,8648 \\ -0,2161 & 1,2969 \end{Vmatrix}$$

și:

$$\|A^{-1}\|_\infty = \max\{10^8 \cdot (0,1441 + 0,8648), \quad 10^8 \cdot (0,2161 + 1,2969)\} = $$
$$= 10^8 \cdot 1,5130$$

iar:

$$\|A\|_\infty = 2,1617.$$

Deci

$$\text{cond}_\infty(A) = 2,1617 \cdot 1,5130 \cdot 10^8 \approx 3,3 \cdot 10^8. \quad (32)$$

Relația (32) ne arată că sistemul considerat este foarte rău condiționat, deci ar trebui să lucrăm în coeficienți și în b cu o precizie mai mare de opt zecimale.

Observație importantă. Chiar dacă elementele matricei A și a vectorului b sînt date exact, totuși din reprezentarea în calculator (virgulă mobilă) rezultă o matrice rotunjită $\bar{A}$ cu elementele $\bar{a}_{ij}$ unde:

$$\frac{|\bar{a}_{ij} - a_{ij}|}{|a_{ij}|} \leqslant u,$$

u reprezentînd unitatea rotunjită în reprezentarea numerelor în virgulă mobilă pe calculator. În cele mai multe mașini de calcul „u" se află între 10^{-15} și 10^{-6}.

§ 7. Metode iterative pentru rezolvarea sistemelor liniare. Metoda Gáuss-Seidel

Să considerăm următorul sistem de două ecuații cu două necunoscute

$$\begin{cases} 3x + y = 4 \\ x - 3y = -2. \end{cases} \quad (34)$$

Soluția sistemului este $x = 1$, $y = 1$.

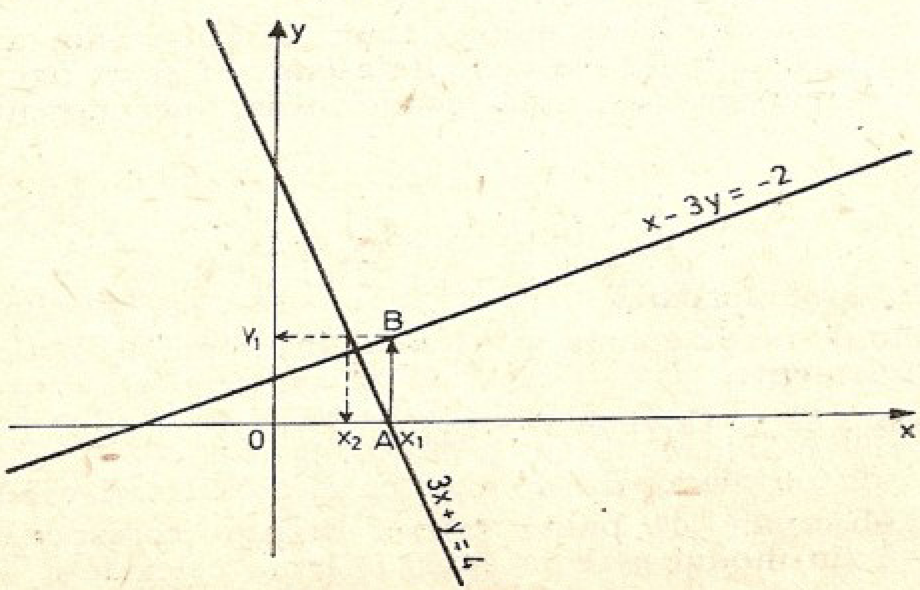


Fig. II.3

Scoatem din prima ecuație pe x și din a doua pe y. Obținem:

$$\begin{cases} x = \frac{1}{3}(4 - y) \\ \\ y = -\frac{1}{3}(-2 - x). \end{cases} \tag{35}$$

Vom considera următorul proces iterativ:

$$\begin{cases} x^{(k)} = \frac{1}{3}(4 - y^{(k-1)}) \\ \\ y^{(k)} = -\frac{1}{3}(-2 - x^{(k)}) \end{cases} \tag{36}$$

plecînd de la o aproximație inițială (de regulă se pleacă de la $x = 0$, $y = 0$).

Deci fie $y^{(0)} = 0$ prima aproximație, atunci din prima ecuație a lui (36) obținem:

$$x^{(1)} = \frac{1}{3} \cdot 4 = \frac{4}{3} \text{ (vezi graficul din figura II.3).}$$

Apoi din a doua ecuație:

$$y^{(1)} = -\frac{1}{3}\left(-2 - \frac{4}{3}\right) = \frac{10}{9} \text{ (figura II.3).}$$

Continuînd procesul de calcul din (36) scoatem:

$$x^{(2)} = \frac{1}{3}\left(4 - \frac{10}{9}\right) = \frac{26}{27},$$

$$y^{(2)} = -\frac{1}{3}\left(-2 - \frac{26}{27}\right) = \frac{80}{81}.$$

Observăm că deși am făcut numai două iterații ne-am apropiat destul de mult de soluția sistemului (se vede de asemenea și pe figura 3).

Dacă scriem matricea asociată coeficienților necunoscutelor sistemului (34):

$$\begin{Vmatrix} 3 & 1 \\ 1 & -3 \end{Vmatrix}$$

observăm următoarele lucruri:

a) elementele de pe diagonala principală sînt nenule și mai mari în modul decît celelalte elemente.

b) $|a_{11}| = |a_{22}| = 3 \cdot |a_{12}| = 3 \cdot |a_{21}|$

deci elementele de pe diagonală domină cu raportul trei celelalte elemente.

Geometric, observăm că panta primei drepte (corespunzătoare primei ecuații) este -3 (în modul mai mare ca 1) iar a celei de a doua este mai mică decît 1.

Acum, să vedem ce se întîmplă dacă permutăm ecuațiile între ele. Deci:

$$\begin{cases} x = 3y - 2 \\ y = \;\; 4 - 3x. \end{cases} \tag{37}$$

Pornind cu aceeași aproximație inițială ($x = 0$, $y = 0$), obținem

Iterația	x	y
0	0	0
1	—2	10
2	28	—80
3	—242	730
. . . . . .	. . . . . .	. . . . . .

(38)

După cum putem observa, procesul iterativ nu converge (figura II. 4).

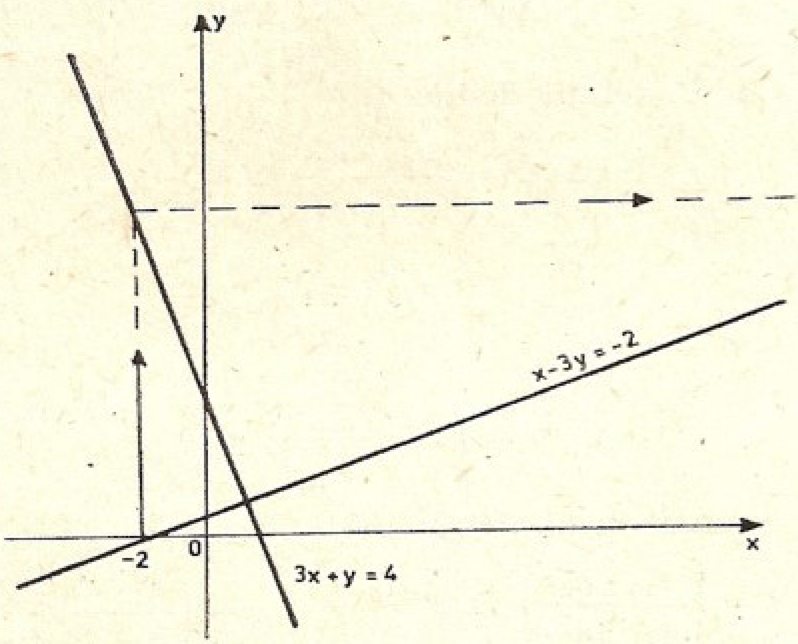


Fig. II. 4

Ce s-a întîmplat?

Panta primei drepte este mai mică decît 1 iar a celei de a doua este mai mare decît 1 (în modul).

Urmează că pentru stabilirea convergenței va trebui să studiem matricea coeficienților necunoscutelor sistemului.

Considerăm pentru aceasta forma generală a unui sistem de două ecuații cu două necunoscute:

$$\begin{cases} a_{11}x + a_{12}y = b_1 \\ a_{21}x + a_{22}y = b_2. \end{cases} \tag{39}$$

Procesul iterativ este definit prin:

$$
\begin{aligned}
x^{(k)} &= \frac{1}{a_{11}} (b_1 - a_{12} y^{(k-1)}) \\
y^{(k)} &= \frac{1}{a_{22}} (b_2 - a_{21} x^{(k)}).
\end{aligned} \tag{40}
$$

Notăm:

$$
\Delta x^{(k)} = x - x^{(k)}
$$

$$
\Delta y^{(k)} = y - y^{(k)}.
$$

Din prima ecuație (39) și prima ecuație din (40) deducem:

$$
\Delta x^{(k)} = - \frac{a_{12}}{a_{11}} \Delta y^{(k-1)}
$$

și similar din celelalte două ecuații:

$$
\Delta y^{(k)} = - \frac{a_{21}}{a_{22}} \Delta x^{(k)}.
$$

Din ultimele două relații deducem:

$$
\Delta y^{(k)} = \frac{a_{12} \cdot a_{21}}{a_{11} \cdot a_{22}} \Delta y^{(k-1)}. \tag{41}
$$

Analog,

$$
\Delta y^{(k-1)} = \frac{a_{12} \cdot a_{21}}{a_{11} \cdot a_{22}} \Delta y^{(k-2)}.
$$

Deci:

$$
\Delta y^{(k)} = \left(\frac{a_{12} \cdot a_{21}}{a_{11} \cdot a_{22}} \right)^2 \Delta y^{(k-2)}.
$$

Continuînd raționamentul, obținem:

$$
\Delta y^{(k)} = \left(\frac{a_{12} \cdot a_{21}}{a_{11} \cdot a_{22}} \right)^3 \Delta y^{(k-3)} = \ldots = \left(\frac{a_{12} \cdot a_{21}}{a_{11} \cdot a_{22}} \right)^k \Delta y^{(0)}. \tag{42}
$$

La fel se obține:

$$
\Delta x^{(k)} = \left(\frac{a_{12} \cdot a_{21}}{a_{11} \cdot a_{22}} \right)^k \Delta x^{(0)}. \tag{43}
$$

Relațiile (42) și (43) ne spun că șirul de iterate $\{x^{(k)}\}$ converge către x (prima componentă a soluției sistemului (39)) și $\{y^{(k)}\}$ converge către y (a doua componentă a soluției sistemului (39)) dacă:

$$
\left| \frac{a_{12} \cdot a_{21}}{a_{11} \cdot a_{22}} \right| < 1. \tag{44}
$$

Condiția (44) este îndeplinită dacă, de exemplu:

$$
\begin{cases} |a_{11}| > |a_{12}| \\ |a_{22}| \geqslant |a_{21}| \end{cases} \tag{45}
$$

sau:

$$\begin{cases} |a_{11}| \geqslant |a_{12}| \\ |a_{22}| > |a_{21}|. \end{cases} \tag{46}$$

Cu alte cuvinte am ajuns la concluzia că diagonala matricei asociate sistemului domină celelalte elemente.

Vom descrie în continuare **metoda Gauss-Seidel** pentru sisteme liniare de trei ecuații cu trei necunoscute.

Fie sistemul:

$$\begin{cases} a_{11}x_1 + a_{12}\,x_2 + a_{13}\,x_3 = b_1 \\ a_{21}x_1 + a_{22}\,x_2 + a_{23}\,x_3 = b_2 \\ a_{31}x_1 + a_{32}\,x_2 + a_{33}\,x_3 = b_3. \end{cases} \tag{47}$$

Presupunem că $a_{11} \neq 0$, $a_{22} \neq 0$, $a_{33} \neq 0$. Urmează că sistemul (47) se poate scrie sub forma:

$$\begin{cases} x_1 = \dfrac{1}{a_{11}}\,(b_1 - a_{12}x_2 - a_{13}x_3) \\ x_2 = \dfrac{1}{a_{22}}\,(b_2 - a_{21}x_1 - a_{23}x_3) \\ x_3 = \dfrac{1}{a_{33}}\,(b_3 - a_{31}x_1 - a_{32}x_2). \end{cases} \tag{48}$$

Vom nota prima aproximație cu $x_1^{(0)}$, $x_2^{(0)}$, $x_3^{(0)}$.

Calculăm noile componente după cum urmează:

$$\begin{aligned} x_1^{(1)} &= \frac{1}{a_{11}}\,(b_1 - a_{12}x_2^{(0)} - a_{13}x_3^{(0)}) \\ x_2^{(1)} &= \frac{1}{a_{22}}\,(b_2 - a_{21}x_1^{(1)} - a_{23}x_3^{(0)}) \\ x_3^{(1)} &= \frac{1}{a_{33}}\,(b_3 - a_{31}x_1^{(1)} - a_{32}x_2^{(1)}). \end{aligned} \tag{49}$$

Astfel s-a încheiat prima iterație. Acum punctul $x_1^{(1)}$, $x_2^{(1)}$, $x_3^{(1)}$ devine aproximație inițială pentru iterația a doua (ia locul, în memoria calculatorului, punctului $x_1^{(0)}$, $x_2^{(0)}$, $x_3^{(0)}$).

În general algoritmul este dat de:

$$\begin{aligned} x_1^{(k)} &= \frac{1}{a_{11}}\,(b_1 - a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)}) \\ x_2^{(k)} &= \frac{1}{a_{22}}\,(b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k-1)}) \\ x_3^{(k)} &= \frac{1}{a_{33}}\,(b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)}). \end{aligned} \tag{50}$$

Putem trece acum uşor la algoritmul Gauss-Seidel pentru un sistem de n ecuaţii cu n necunoscute.

Dacă matricea coeficienţilor este $A = \|a_{ij}\|_{\substack{i=\overline{1,n}\\ j=\overline{1,n}}}$

iar vectorul termenilor liberi este: $b = \begin{Vmatrix} b_1 \\ \vdots \\ b_n \end{Vmatrix}$ atunci:

$$x_i^{(k+1)} = \frac{-\sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_i}{a_{ii}} \tag{60}$$

pentru $i = 1, 2, 3, \ldots, n$.

Ca primă aproximaţie se consideră, de obicei, un vector n-dimensional cu toate componentele nule.

Calculul se opreşte cînd:

$$\max_{i=1,\ldots,n} | x_i^{(k)} - x_i^{(k-1)} | < \varepsilon. \tag{61}$$

Altfel spus, cînd componentele vectorului soluţie la două iteraţii succesive devin foarte apropiate înseamnă că ne-am apropiat de soluţie.

Din discuţiile anterioare se poate deduce că metoda descrisă converge cînd:

$$|a_{ii}| \geqslant |a_{i1}| + |a_{i2}| + \ldots + |a_{in}| \tag{62}$$

pentru orice $i = 1, 2, 3, \ldots, n$ şi pentru cel puţin un i avem inegalitate strictă. Deci:

$$|a_{ii}| > |a_{i1}| + |a_{i2}| + \ldots + |a_{in}|. \tag{63}$$

De exemplu pentru matricea $n \times n$

$$A = \begin{Vmatrix} 2 & -1 & & & & & \mathbf{0} \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & -1 & 2 & -1 & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & \cdot \\ & & & & & \cdot & \cdot \quad -1 \\ \mathbf{0} & & & & & -1 & 2 \end{Vmatrix} \tag{64}$$

metoda Gauss-Seidel converge.

Exerciţiu. — *Explicaţi de ce!*

Dăm în continuare, organigrama (fig. II. 5) şi programul de calcul pentru metoda Gauss-Seidel. Urmează de asemenea un exemplu de calcul.

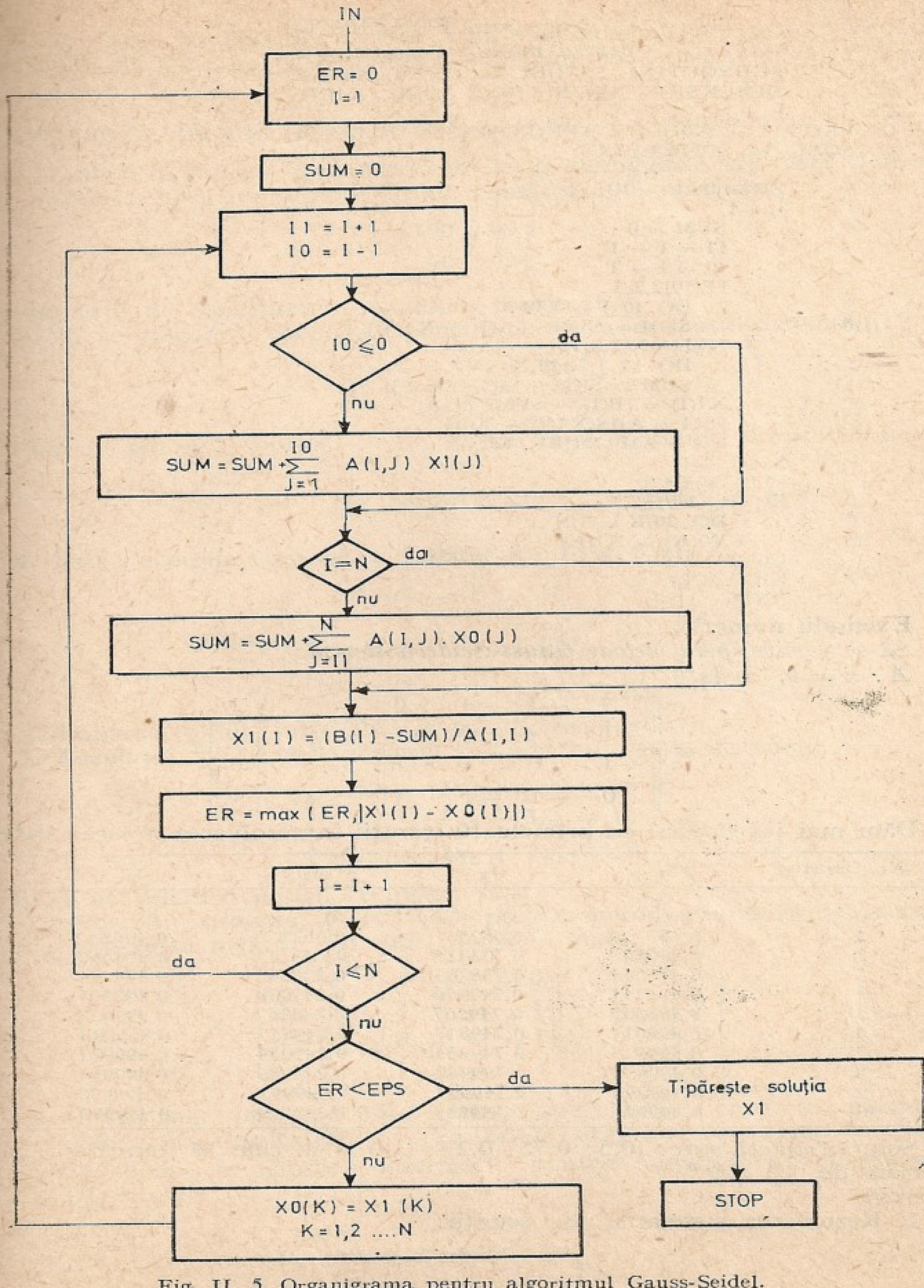


Fig. II. 5. Organigrama pentru algoritmul Gauss-Seidel.

Programul FORTRAN
pentru metoda Gauss-Seidel

```
      SUBROUTINE SEIDEL (A, B, X0, X1,N,EPS)
      DIMENSION A(N, N), B(N), X0(N), X1(N)
C
C  VECTORUL X0 ESTE APROXIMAŢIA INIŢIALĂ. ÎN X1(N) SE OBŢINE
C  SOLUŢIA SISTEMULUI
C
 8    ER = 0
      I = 1
 4          SUM = 0
            I1 = I + 1
            I0 = I − 1
            IF(I0)2,2,3
 3              DO 10 J = 1,I0
10              SUM = SUM + A(I, J).X1(J)
            IF(I1−N)2,2,12
 2              DO 11 J = I1,N
11              SUM = SUM + A(I,J).X0(J)
12          X1(I) = (B(I) − SUM)/A(I,I)
            ER1 = ABS(X1(I) − X0(I))
            ER = AMAX1(ER, ER1)
      I = I + 1
      IF(I−N)4,4,5
 5    IF(ER−EPS)7,7,6
 6          DO 20 K = 1,N
20          X0(K) = X1(K)
      GO TO 8
 7    RETURN
      END
```

Exemplu numeric.

Să se *rezolve, prin metoda Gauss-Seidel sistemul:*

$A \cdot x = b$, *unde:*

$$A = \begin{Vmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{Vmatrix}, \quad b = \begin{Vmatrix} 1 \\ 2 \\ 0 \\ 1 \end{Vmatrix}.$$

Dăm mai jos tabelul cu primele 10 iteraţii în rezolvarea acestui sistem:

Nr. iteraţiei	x_1	x_2	x_3	x_4
0	0	0	0	0
1	0,25	0,5625	0,0625	0,40625
2	0,40625	0,703125	0,203125	0,476563
3	0,476563	0,738281	0,238281	0,494141
4	0,494141	0,747070	0,247070	0,498535
5	0,498535	0,749267	0,249267	0,499634
6	0,499634	0,749817	0,249817	0,499908
7	0,499908	0,749954	0,249954	0,499971
8	0,499971	0,749989	0,249989	0,499994
9	0,499994	0,749997	0,249997	0,499998
10	0,499998	0,749999	0,249999	0,499999

Soluţia exactă este: (0,5; 0,75; 0,25; 0,5) şi în cele 10 iteraţii s-a atins precizia de $1{,}5 \cdot 10^{-6}$

Exerciţii

1. Rezolvarea sistemului de ecuaţii:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \end{cases}$$

revine la a găsi punctul de intersecție al celor două drepte date de ecuațiile sistemului. Să se arate că unghiul făcut de cele două drepte verifică inegalitatea:

$|\operatorname{ctg} \alpha| \leqslant \frac{1}{2} \operatorname{cond}_E(A)$ (se presupune matricea A nesingulară).

2. Arătați că printre toate matricele 2×2, nesingulare, ale căror elemente sînt întregi nenegativi și nedepășind numărul 100, matricea

$$\begin{Vmatrix} 100 & 99 \\ 99 & 98 \end{Vmatrix}$$

are numărul de condiționare euclidian maximal.

3. Să se calculeze $\operatorname{cond}_\infty(A)$ pentru matricea asociată sistemului:

$$\begin{cases} 5x - 3{,}31y = 1{,}69 \\ 6x - 3{,}97y = 2{,}03. \end{cases}$$

Cum se schimbă soluția acestui sistem dacă membrul drept devine:

$$b = \begin{Vmatrix} 1{,}7 \\ 2 \end{Vmatrix}$$ (deci s-a schimbat puțin membrul drept).

4. Este nevoie la rezolvarea sistemului: $Ax = b$ unde:

$$A = \begin{Vmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{Vmatrix}; \quad b = \begin{Vmatrix} 1 \\ 1 \\ 1 \\ 1 \end{Vmatrix}$$

de schimbarea liniilor între ele? Explicați răspunsul.

5. Rezolvați, pe calculator, prin metoda lui Gauss sistemul:

$$\begin{cases} 9x_1 - 2x_2 + x_3 = 50 \\ x_1 + 5x_2 - 3x_3 = 18 \\ -2x_1 + 2x_2 + 7x_3 = 19 \end{cases}$$

utilizați în calcul 6 poziții zecimale.

6. Scrieți un program FORTRAN pentru metoda Gauss-Seidel și rezolvați prin această metodă sistemul: $Ax = b$, unde:

$$A = \begin{Vmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{Vmatrix}; \quad b = \begin{Vmatrix} 1 \\ 2 \\ 0 \\ 1 \end{Vmatrix}.$$

a) În cîți pași obțineți o bună aproximație? (lucrați cu 5 cifre zecimale).

b) Comparați rezultatele cu rezultatele obținute prin metoda Gauss.

7. Să se dea o margine inferioară pentru numărul de condiționare $\operatorname{cond}_\infty(A)$ unde:

$$A = \begin{Vmatrix} 1 & -1 & 1 \\ -1 & \varepsilon & \varepsilon \\ 1 & \varepsilon & \varepsilon \end{Vmatrix}, \qquad \varepsilon \neq 0.$$

8. Evaluați perturbarea eventuală a soluțiilor sistemului:

$$\begin{cases} x - 2y = -1 \\ -2x + 4{,}01y = 2 \end{cases}$$

dacă schimbăm componentele membrului drept cu 0,01. Găsiți soluția sistemului cu aceeași matrice a coeficienților și cu membrul drept dat de:

$$b = \begin{Vmatrix} -1 \\ 2{,}01 \end{Vmatrix}.$$

9. Arătați că schimbările de linii sau de coloane într-o matrice A nu schimbă numerele de condiționare $\text{cond}_E(A)$; $\text{cond}_\infty(A)$.

Exerciții facultative (cap. II)

Se dă matricea (de dimensiune $n \times n$)

$$A = \begin{Vmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & -1 \\ 0 & \cdots & \cdots & -1 & 2 \end{Vmatrix}$$

a) arătați că determinantul lui A este $n + 1$;

b) rezolvați ecuația de gradul n:

$$\begin{vmatrix} 2-\lambda & -1 & 0 & \cdots & 0 \\ -1 & 2-\lambda & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & 2-\lambda & -1 \\ 0 & \cdots & \cdots & -1 & 2-\lambda \end{vmatrix} = 0$$

utilizînd substituția $\lambda = 2\cos\theta$.

c) Notînd cea mai mare rădăcină a ecuației cu $\lambda_n^{(n)}$ și cea mai mică cu $\lambda_1^{(n)}$ calculați:

$$\lim_{n\to\infty} \frac{\lambda_n^{(n)}}{\lambda_1^{(n)}}.$$

d) Arătați că inversa matricei A este dată de:

$$\lambda_{\mu\nu} = \begin{cases} \dfrac{(n+1-\mu)\nu}{n+1}, & \nu \leqslant \mu \\[2ex] \dfrac{\mu(n+1-\nu)}{n+1}, & \nu \geqslant \mu \end{cases}$$

e) Aplicați algoritmul lui Gauss pentru a inversa matricea A, de dimensiuni 10×10, utilizînd în membrul drept matricea unitate.

Capitolul III

REZOLVAREA ECUAȚIILOR NELINIARE

§ 1. Introducere. Metode iterative

Dacă $f(x)$ este o funcție reală de variabilă reală, atunci soluțiile ecuației $f(x) = 0$ în general nu pot fi exprimate sub o formă convenabilă (o formulă de exemplu) și în consecință, pentru obținerea lor, se adoptă metode iterative.

Pentru a ne face o idee despre ceea ce înseamnă metodă iterativă, în acest cadru, vom considera ecuația:

$$x = F(x) \tag{1}$$

unde $F(x)$ este o funcție (de regulă derivabilă de mai multe ori) ale cărei valori le putem calcula pentru x aparținînd unui interval.

Se pleacă de la un punct x_0, **aproximație inițială,** și calculăm șirul:

$$x_1 = F(x_0)\,;\;\; x_2 = F(x_1)\,;\;\; x_3 = F(x_2), \ldots \tag{2}$$

Fiecare etapă din calcul:

$$x_{k+1} = F(x_k) \tag{3}$$

se numește **iterație.**

Dacă șirul $\{x_n\}$ converge către o valoare limită α atunci din continuitatea funcției $F(x)$ rezultă că:

$$\lim_{n\to\infty} F(x_n) = F(\alpha)$$

deci, trecînd la limită în (3) obținem:

$$\alpha = F(\alpha) \tag{4}$$

adică α satisface ecuația $x = F(x)$.

O interpretare geometrică a acestui calcul iterativ este dată în figura III.1.

Se vede că, plecînd din x_0 obținem $F(x_0)$. Apoi ducem o paralelă la axa Ox prin punctul $(x_0, F(x_0))$ pînă intersectează prima bisectoare $y = x$, coborîm din acel punct perpendiculara pe OX și obținem pe $x_1(x_1 = F(x_0))$, apoi $x_2 = F(x_1)$ și procedeul continuă.

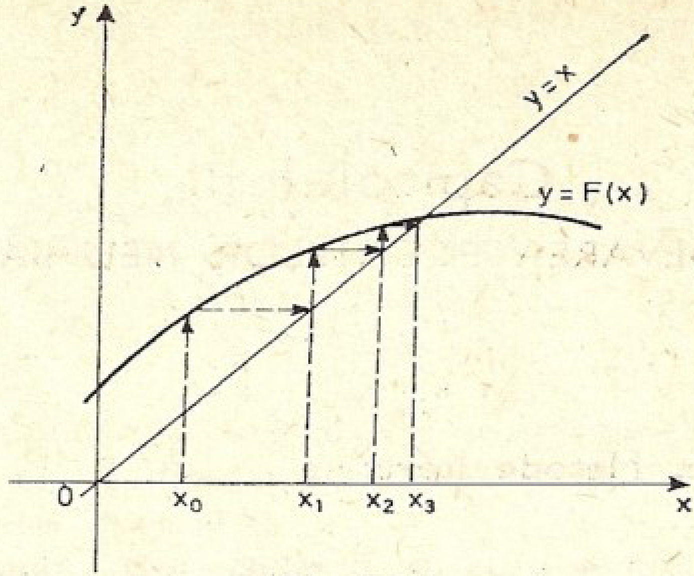


Fig. III. 1

Dacă urmărim graficul funcţiei din figura III.1 observăm că în apropierea soluţiei pantele tangentelor la graficul funcţiei $y = F(x)$ devin foarte mici (pozitive).

Acest fapt ne conduce la a considera comportarea primei derivate în vecinătatea soluţiei.

Utilizînd *teorema de medie a lui* **Lagrange** obţinem:

$$\frac{x_{n+1} - x_n}{x_n - x_{n-1}} = \frac{F(x_n) - F(x_{n-1})}{x_n - x_{n-1}} = F'(\zeta_n) \tag{5}$$

unde ζ_n se află între x_{n-1} şi x_n.

În (5) am utilizat şi relaţia de iterare $x_{n+1} = F(x_n)$.

Urmează că, dacă $|F'(x)| < 1$ pentru toţi x într-o vecinătate a rădăcinii care conţine pe x_0 şi x_1, atunci avem convergenţa asigurată.

Acum prezentăm riguros demonstraţia următoare:

Teoremă. Dacă ecuaţia $x = F(x)$ are o rădăcină „α" şi $F'(x)$ există în intervalul $J = \{x;\ |x - \alpha| \leqslant \rho\}$ şi satisface condiţia:

$|F'(x)| \leqslant m < 1$ **atunci pentru orice** $x_0 \in J$ (deci orice aproximaţie iniţială în intervalul J) **avem:**

a) $x_n \in J$, $n = 0, 1, 2, \ldots,$

b) $\lim\limits_{n\to\infty} x_n = \alpha,$

c) **Singura soluţie a ecuaţiei în intervalul J este α.**

Demonstraţie. (a) Demonstrăm prin inducţie. Fie $x_0 \in J$, presupunem că $x_{n-1} \in J$, atunci arătăm că şi $x_n \in J$.

Într-adevăr, conform teoremei lui Lagrange avem:

$$x_n - \alpha = F(x_{n-1}) - F(\alpha) = F'(\zeta_n)\,(x_{n-1} - \alpha)$$

unde $\zeta_n \in J$.

De unde:

$$|x_n - \alpha| = F'(\zeta_n)(x_{n-1} - \alpha)| \leqslant m \cdot |x_{n-1} - \alpha| \leqslant m\rho$$

deoarece $x_{n-1} \in J$.

Urmează că $x_n \in J$ şi (a) este demonstrată.

(b) Avem:

$$|x_n - \alpha| \leqslant m\,|x_{n-1} - \alpha|$$

ontinuînd raţionamentul obţinem:

$$|x_n - \alpha| \leqslant m \cdot |F(x_{n-2}) - F(\alpha)| = m \cdot |F'(\bar{\zeta}_n)(x_{n-2} - \alpha)| \leqslant m^2 \cdot |x_{n-2} - \alpha|,$$

$$(\bar{\zeta}_n \in J).$$

De unde, recursiv se obţine:

$$|x_n - \alpha| \leqslant m^n\,|x_0 - \alpha|.$$

Cum $0 < m < 1$ avem $\lim\limits_{n \to \infty} m^n = 0$.

Trecînd la limită în ambii termeni ai inegalităţii precedente obţinem:

$$\lim_{n \to \infty} |x_n - \alpha| = 0, \text{ deci } \lim_{n \to \infty} x_n = \alpha.$$

(c) Presupunem că ecuaţia $x = F(x)$ ar avea în J şi rădăcina $\beta \neq \alpha$. Atunci:

$$\beta = F(\beta) \text{ deci,}$$

$$\alpha - \beta = F(\alpha) - F(\beta) = F'(\eta)(\alpha - \beta); \quad \eta \in J$$

de unde:

$$|\alpha - \beta| \leqslant m \cdot |\alpha - \beta| < |\alpha - \beta|$$

deci o contradicţie şi implicaţia (c) este adevărată.

§ 2. O metodă rapidă pentru extragerea rădăcinii pătrate

Ecuaţia $x^2 = c$ poate fi scrisă şi sub forma $x = F(x)$ unde:

$$F(x) = \frac{1}{2}\left(x + \frac{c}{x}\right), \quad c > 0. \tag{6}$$

Soluţia ecuaţiei $x = F(x)$ este $\alpha = \sqrt{c}$ (valoarea limită a algoritmului $x_{n+1} = F(x_n)$) (fig. III.2).

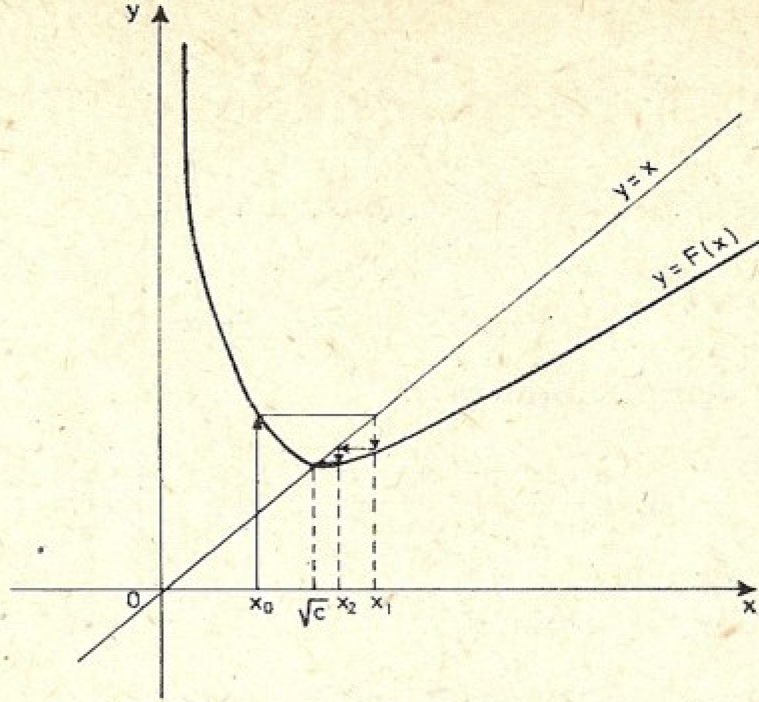


Fig. III. 2

De data aceasta algoritmul se reduce la șirul inductiv:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right). \tag{7}$$

Dăm în figura III.3 organigrama apoi programul FORTRAN pentru calculul rădăcinii pătrate a unui număr.

De asemenea dăm tabelul cu iterațiile în calculul rădăcinii pătrate a lui 2 plecînd de la $x_0 = 1{,}5$.

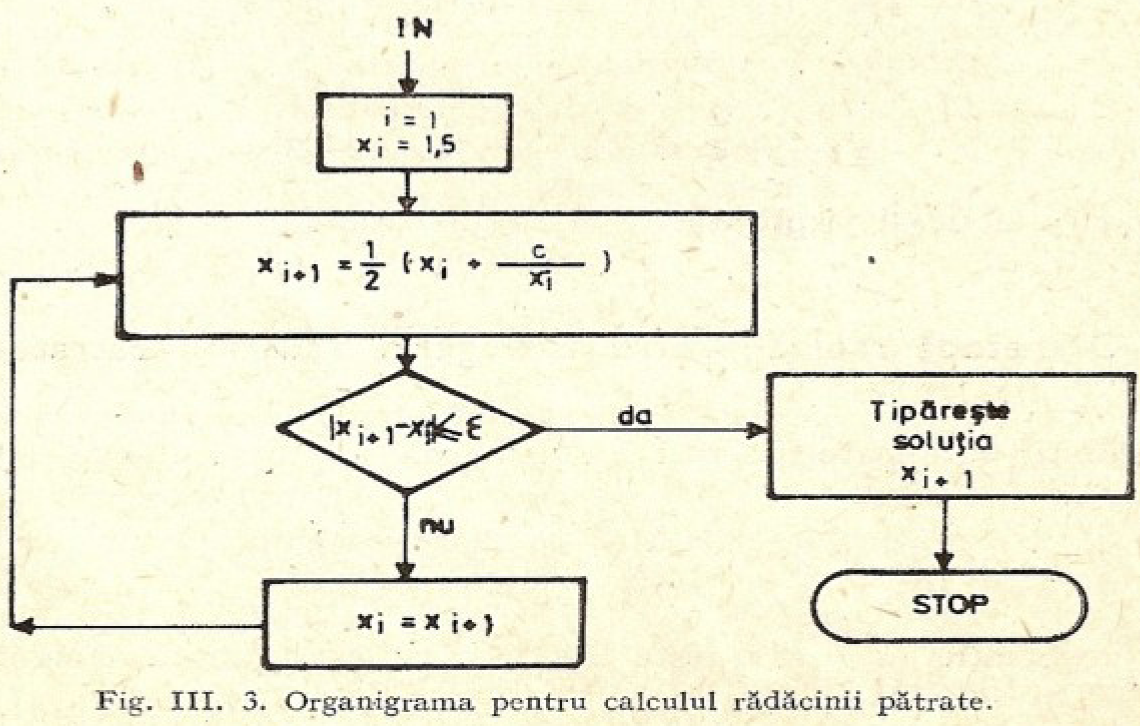


Fig. III. 3. Organigrama pentru calculul rădăcinii pătrate.

Urmează programul FORTRAN pentru calculul rădăcinii pătrate:

```
 1  FORMAT (6E 12.5)
    READ (105,1) C,EPS
    XI = 1·5
12  XI1 = (XI + C/XI)/2.0
    IF(ABS(XI1-XI) - EPS) 10, 10, 11
11  XI = XI1
    GO TO 12
10  WRITE (108,2) C, XI1
 2  FORMAT (8X, 'RĂDĂCINA PĂTRATĂ A NUMĂRULUI',
    1E 12.5, 'ESTE', E.12.5)
    STOP
    END
```

Dăm în continuare tabelul cu primele trei iterații în calculul rădăcinii pătrate din numărul 2:

Nr. iterației	x_i	$x_{i+1} = \frac{1}{2}\left(x_i + \frac{2}{x_i}\right)$
0	1,5	—
1	1,5	1,416667
2	1,416667	1,414216
3	1,414216	1,414214
	$(\sqrt{2} = 1{,}414213562\ldots)$	

§ 3. Analiza erorilor

Calculul valorilor funcției $F(x)$ este afectat de erori (de exemplu de erori de rotunjire).

Vom face, în cele ce urmează, o evaluare a erorii după un număr finit de iterații.

Să notăm cu $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_n, \ldots$ șirul valorilor calculate. De asemenea, notăm cu δ_n eroarea din calculul lui $F(\bar{x}_n)$.

Avem:

$$\bar{x}_{n+1} = F(\bar{x}_n) + \delta_n, \quad n = 0, 1, 2, \ldots, n. \tag{8}$$

Fie α rădăcina ecuației $x = F(x)$, deci:

$$\alpha = F(\alpha). \tag{9}$$

Scăzînd relația (9) din (8) obținem:

$$\bar{x}_{n+1} - \alpha = F(\bar{x}_n) - F(\alpha) + \delta_n \tag{10}$$

Aplicînd teorema lui Lagrange, obținem:

$$\bar{x}_{n+1} - \alpha = F'(\zeta_n)\,(\bar{x}_n - \alpha) + \delta_n \tag{11}$$

unde ζ_n aparține intervalului determinat de $\bar{x}_n$ și de α.

Relația (11) se mai scrie:

$$(1 - F'(\zeta_n))(\bar{x}_{n+1} - \alpha) = F'(\zeta_n)(\bar{x}_n - \bar{x}_{n+1}) + \delta_n. \qquad (12)$$

Presupunem că $|F'(\zeta_n)| \leqslant m < 1$ și $|\delta_n| \leqslant \delta$ atunci (12) ne dă:

$$|\bar{x}_{n+1} - \alpha| \leqslant \frac{m}{1-m}|\bar{x}_{n+1} - \bar{x}_n| + \frac{1}{1-m}\delta. \qquad (13)$$

Primul termen al membrului drept estimează eroarea de trunchiere iar al doilea termen estimează eroarea de calcul.

Eroarea de calcul $\frac{\delta}{1-m}$ depinde de eroarea δ_n numai din ultima iterație și deci $\bar{x}_n$ poate fi considerată o cantitate exactă cînd putem estima pe δ_n.

Acest lucru ne conduce la următoarea concluzie: Nu este necesar să calculăm cu mare precizie primele iterate deoarece rotunjirea în aceste iterații nu are efect asupra rezultatului final.

Se spune că metodele iterative se autocorectează.

Studiind relația (13) se observă că pentru „n" suficient de mare în diferența $|\bar{x}_{n+1} - \alpha|$ rolul dominant îl are eroarea de calcul $\frac{\delta}{1-m}$. Așa că, dacă după un „n" suficient de mare vom continua iterațiile, vom observa că diferența $|\bar{x}_{n+1} - \alpha|$ se comportă ca o mărime de ordinul $\frac{\delta}{1-m}$.

Exemplu. *Pentru calculul lui $\sqrt{2}$ plecînd de la $x_0 = 1,5$ obținem:*

$$\bar{x}_1 = \frac{1}{2}\left(1,5 + \frac{2}{1,5}\right) = 1,415 \text{ (rotunjit la 3 zecimale)},$$

$$\bar{x}_2 = \frac{1}{2}\left(x_1 + \frac{2}{x_1}\right) = 1,4142 \text{ (rotunjit la 4 zecimale)},$$

$$\bar{x}_3 = \frac{1}{2}\left(x_2 + \frac{2}{x_2}\right) = 1,41422 \text{ (rotunjit la 5 zecimale)}.$$

$$\bar{x}_4 = \frac{1}{2}\left(x_3 + \frac{2}{x_3}\right) = 1,41422 \text{ (rotunjit la 5 zecimale)}.$$

Pentru $\bar{x}$ aparținînd intervalului generat de $\bar{x}_4$ și $\sqrt{2}$ avem:

$$|\bar{x}_4 - \sqrt{2}| \leqslant \frac{|F'(\zeta_n)|}{1 - |F'(\zeta_n)|}|\bar{x}_4 - \bar{x}_3| + \frac{1}{1 - |F'(\zeta_n)|}\cdot\delta$$

însă:

$$F'(x) = \frac{x^2 - 2}{x^2} \text{ și observăm că}$$

$F'(\bar{x}_4) \geqslant F'(\zeta_n)$ pentru ζ_n cuprins în intervalul $[\bar{x}_4, \sqrt{2}]$.

Deci:

$$|\bar{x}_4 - \sqrt{2}| \leqslant \frac{|F'(\bar{x}_4)|}{1 - |F'(\bar{x}_4)|} |\bar{x}_4 - \bar{x}_3| + \frac{1}{1 - |F'(\bar{x}_4)|} \cdot \delta.$$

Efectuînd calculele avem:

$$F'(\bar{x}_4) = 0{,}00009 \approx 1 \cdot 10^{-4},$$

$$\frac{1}{1 - F'(\bar{x}_4)} \approx 1{,}0001,$$

$$|\bar{x}_4 - \bar{x}_3| \leqslant \frac{1}{2} \cdot 10^{-6}$$

$$\delta \leqslant 1 \cdot 10^{-6} \text{ (precizia în calcul)}$$

și cu aceasta:

$$|\bar{x}_4 - \sqrt{2}| \leqslant 0{,}50005 \cdot 10^{-10} + 1{,}0001 \cdot 10^{-6}.$$

Se observă că după 4 iterații precizia depinde numai de eroarea de calcul:

$$\frac{1}{1 - m} \delta = 1{,}0001 \cdot 10^{-6}.$$

Să presupunem acum că x_n este aproximația de ordin n a rădăcinii simple, a ecuației $f(x) = 0$.

Atunci, din teorema lui Lagrange:

$$f(x_n) = (x_n - \alpha) f'(\zeta_n),$$

unde ζ_n aparține intervalului generat de x_n și α.

Dacă $|f'(x)| \geqslant M_1$ pentru $x \in J$ atunci:

$$|x_n - \alpha| \leqslant \frac{|f(x_n)|}{M_1} . \tag{14}$$

Deci avem o estimare a erorii numai funcție de valorile lui $f(x_n)$ (ceea ce este foarte convenabil, de multe ori, dacă cunoaștem comportarea primei derivate).

Să notăm cu $\bar{f}(x_n)$ valoarea calculată a funcției f în x_n și fie:

$$\bar{f}(x_n) = f(x_n) + \delta(x_n), \quad \delta(x_n) \leqslant \delta.$$

Vom putea deci exprima cît de bine este aproximată soluția în funcție de δ.

Situația cea mai bună ar fi să determinăm un x_n astfel ca $\bar{f}(x_n) = 0$. În acest caz valoarea exactă a funcției $f(x)$ satisface inegalitatea:

$$|f(x_n)| \leqslant \delta.$$

Presupunînd că derivata întîi $f'(x)$ nu variază foarte mult în apropierea soluției $x = \alpha$, din (14) deducem:

$$|x_n - \alpha| \leqslant \frac{\delta}{M_1} \approx \varepsilon_\alpha, \text{ unde } \varepsilon_\alpha = \frac{\delta}{|f'(\alpha)|} . \tag{15}$$

Urmează că cea mai bună margine a erorii pentru orice metodă este ε_α care poartă numele de precizia ce se poate atinge pentru rădăcina α.

Dacă $|f'(\alpha)|$ este mic atunci ε_α este mare și în acest caz convergența șirului $\{x_n\}$ este slabă (spunem că problema este rău condiționată).

Dacă privim figura III. 4,

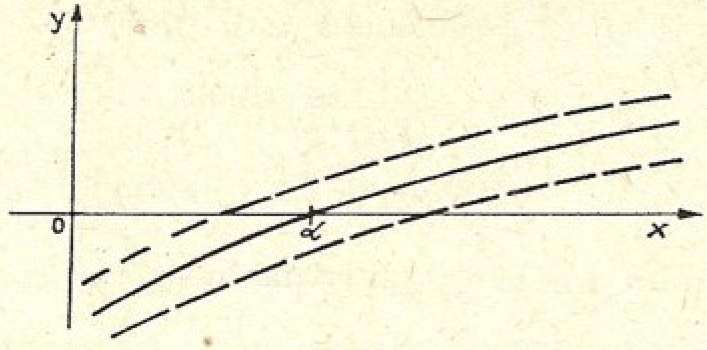


Fig. III. 4

observăm că o mică schimbare în funcție produce o mare variație în evaluarea rădăcinii.

Pentru oprirea practică a procesului iterativ nu se folosește ε_α (nefiind cunoscut) ci următorul criteriu:

$$\begin{cases} |x_{n+1} - x_n| \leqslant |x_n - x_{n-1}| \\ |x_n - x_{n-1}| < \delta. \end{cases} \tag{16}$$

§ 4. Metoda Newton-Raphson

Metoda Newton-Raphson este o metodă de rezolvare a ecuațiilor neliniare bazată pe aproximarea curbei $y = f(x)$ cu tangenta sa în punctul $(x_0, f(x_0))$, apoi următorul punct x_1 se obține intersectînd tangenta la grafic în x_0 cu axa Ox. Procedeul continuă (fig. III. 5). Scriind ecuația tangentei în $(x_0, f(x_0))$ la graficul funcției $y = f(x)$ obținem:
$y - f(x_0) = f'(x_0)\,(x - x_0)$, de unde pentru $y = 0$ avem (presupunînd $f'(x_0) \neq 0$):

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{17}$$

Deci:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{18}$$

Se formează astfel șirul iteratelor $x_1, x_2, x_3, \ldots, x_n, \ldots$ dat de algoritmul:

$$x_{n+1} = x_n + h_n, \text{ unde } h_n = -\frac{f(x_n)}{f'(x_n)}. \tag{19}$$

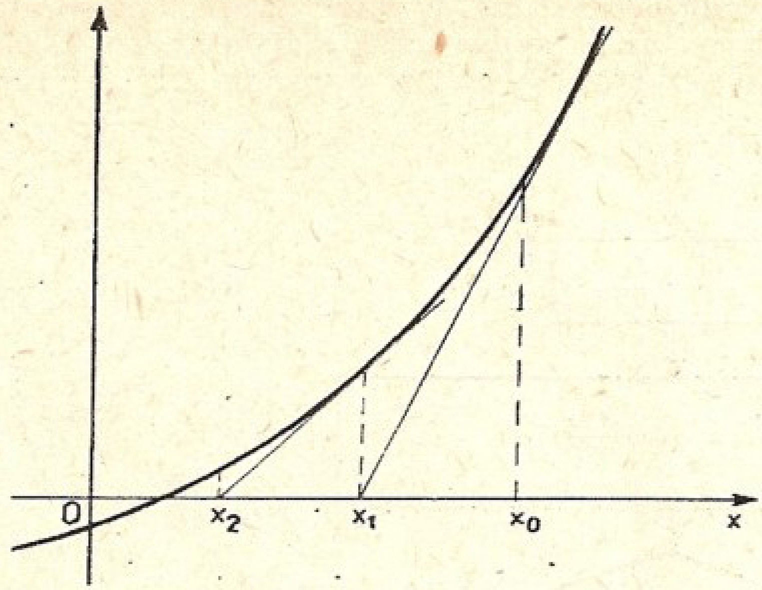


Fig. III. 5

Funcția $f(x)$ se aproximează prin tangenta sa în punctul $(x_n, f(x_n))$ și x_{n+1} este abscisa punctului de intersecție a tangentei cu axa Ox.

Șirul $\{x_n\}$ converge (nu dăm demonstrația aici) către o rădăcină a ecuației $f(x) = 0$, dacă această rădăcină este presupusă simplă și alegerea inițială x_0 este suficient de apropiată de ea.

Dăm mai jos un criteriu de convergență, ușor de aplicat.

Teoremă. — Presupunem că $f'(x) \neq 0$ și $f''(x)$ nu-și schimbă semnul în intervalul $[a, b]$ iar $f(a) \cdot f(b) < 0$. Dacă $\left|\frac{f(a)}{f'(a)}\right| < b - a$, $\left|\frac{f(b)}{f'(b)}\right| < b - a$ atunci metoda Newton-Raphson converge pentru orice alegere inițială $x_0 \in [a, b]$.

Pentru înțelegerea acestei teoreme a se vedea figura III.6.

Exercițiu. *Verificați că teorema precedentă se poate aplica pentru rezolvarea ecuației* $x^3 - 1{,}5 = 0$ *pe intervalul* $[1, 2]$.

Din șirul $\{x_n\}$ dat de relația (19) noi putem calcula, evident numai un număr finit de termeni — de aceea trebuie să spunem cînd oprim procesul de calcul (19) astfel încît să avem o aproximație dorită a rădăcinii căutate.

Criteriul de oprire (vezi programul de calcul) va fi următorul:

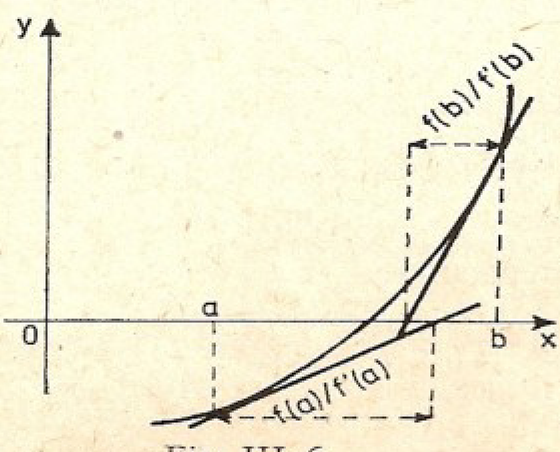


Fig. III. 6

Cînd h_n devine mai mic decît eroarea admisă în soluție atunci procesul de calcul se oprește și x_{n+1} devine soluție aproximativă a ecuației $f(x) = 0$.

Dăm în figura III. 7 organigrama și programul FORTRAN pentru metoda Newton-Raphson:

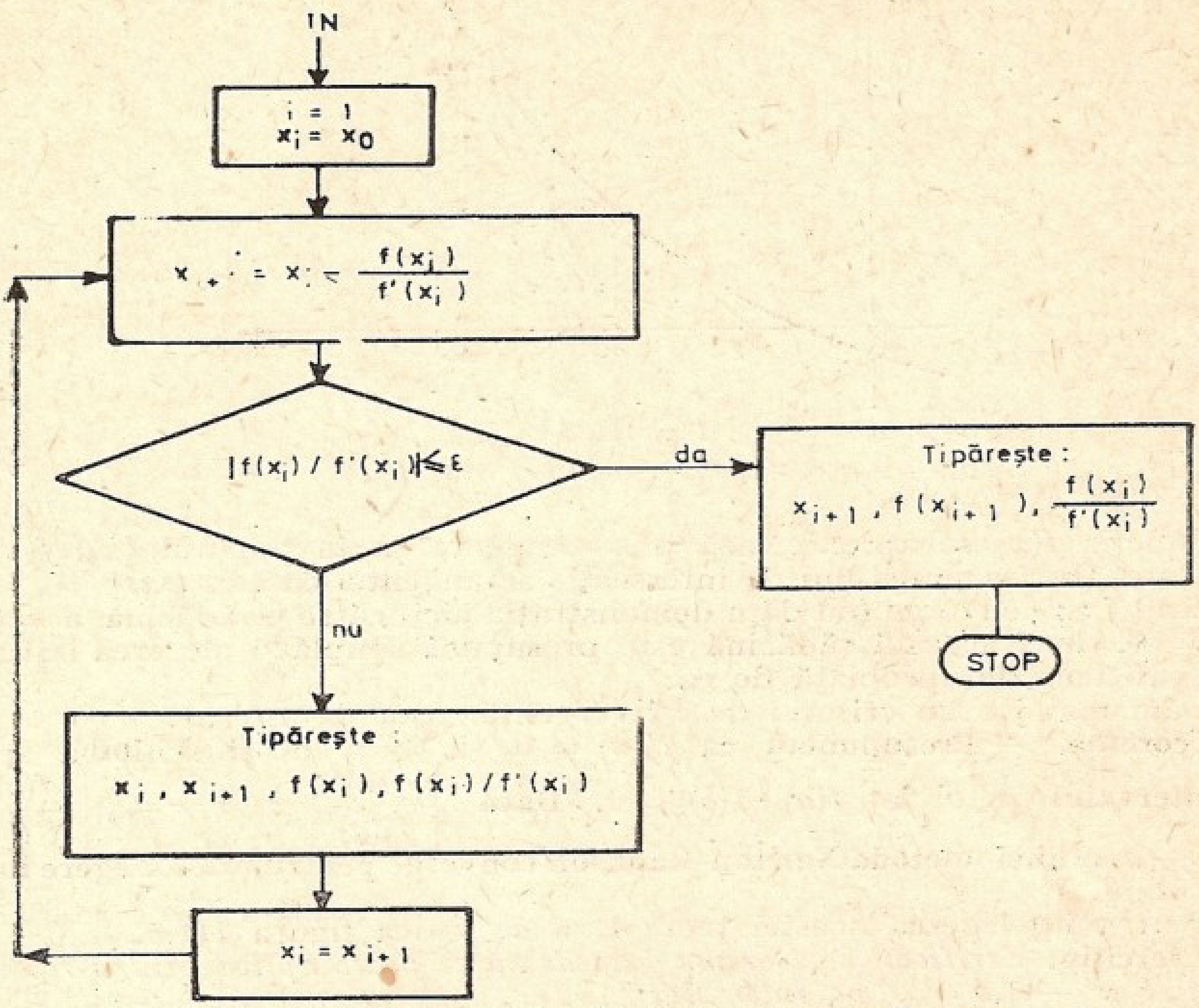


Fig. III. 7. Organigrama pentru metoda Newton-Raphson

Programul FORTRAN

pentru metoda Newton-Raphson,

```
      READ (105, 1) XI, EPS
    1 FORMAT (6E 12.5)
    3 HF = F(XI)
      HI = HF/FD(XI)
      XI1 = XI - HI
      IF(ABS(HI) - EPS) 10, 10, 11
   11 WRITE (108, 2) XI, XI1, HF, HI
    2 FORMAT (8X, 6E 12.5)
      XI = XI1
      GO TO 3
   10 HF = F(XI1)
      WRITE (108, 2) XI1, HF, HI
      STOF
      END
```

```
FUNCTION F(X)
F = ...aici se scrie programul pentru f(x)
RETURN
END
FUNCTION FD(X)
FD = ...aici se scrie programul pentru f'(x)
RETURN
END
```

Programul de mai sus a fost utilizat pentru rezolvarea ecuației:

$$f(x) = \sin x - (x/2)^2 = 0.$$

S-a pornit cu aproximația inițială $x_0 = 1{,}5$. Procesul iterativ a condus la soluție (cu precizia de $1 \cdot 10^{-4}$) în numai patru pași.

Exercițiu. *Rezolvați aceeași problemă pe calculator completînd programul pentru funcțiile $F(X)$ și $FD(X)$. Folosiți în calcul 5 zecimale și porniți cu aproximația $x_0 = 1{,}5$.*

Tipăriți la fiecare iterație următoarele valori:

Nr. iterației	x_n	x_{n+1}	$f(x_n)$	h_n

Exerciții

1. Pentru a intui convergența sau divergența unei metode iterative de rezolvare a ecuației $x = F(x)$ faceți reprezentări grafice pentru cazurile

$$-1 < F'(x) < 0;\ 0 < F'(x) < 1;\ F'(x) > 1;\ F'(x) < -1.$$

2. Explicați cu cuvintele dumneavoastră de ce putem spune că metoda Newton-Raphson este o metodă de „liniarizare" plus iterare.

3. În cîți pași se rezolvă ecuația $x - 3 = 0$ prin metoda Newton-Raphson? Este importantă alegerea inițială?

4. Arătați că ecuația:

$$x^3 - 5x^2 - 45x + 50 = 0 \text{ are o rădăcină cuprinsă între 1 și 2.}$$

Calculați acea rădăcină prin metoda Newton-Raphson alegînd $x_0 = 1$ (calculul cu 6 zecimale corecte).

5. Calculați $\sqrt{2}$ prin metoda lui Newton-Raphson cu 6 zecimale exacte.

6. Calculați $\sqrt[3]{5}$ utilizînd metoda Newton-Raphson:

$$x_{n+1} = \frac{1}{3}\left(2x_n + \frac{5}{x_n^2}\right)$$

(calculele cu 6 zecimale corecte).

7. Găsiți cea mai mică rădăcină pozitivă a ecuației:

$$x - \operatorname{tg} x = 0$$

(cu șase zecimale).

8. În calculul rădăcinii α a ecuației $x = F(x)$ cu o eroare mai mică decît $0{,}5 \cdot 10^{-4}$ s-a obținut:

$$x_4 = 0{,}43789 \text{ și } x_5 = 0{,}43814.$$

Se știe că $|F'(x)| \leqslant 0,4$. Ce număr de iterații sînt necesare pentru a fi siguri că am atins acuratețea cerută?

9. Scrieți un program de calcul (subrutină) pentru rezolvarea ecuației $x = F(x)$, iterativ, $x_{n+1} = F(x_n)$, cu x_0 aproximarea inițială și oprirea calculului dată de:

$$\frac{m \cdot |x_{n+1} - x_n|}{1 - m} < \varepsilon$$

unde m este dat de $\frac{|x_{n+1} - x_n|}{|x_n - x_{n-1}|}$.

10. Aplicați programul precedent la găsirea rădăcinilor ecuației

$$3x - \cos x = 0.$$

11.a) Rădăcina reală a ecuației $x^3 = x + 4$ poate fi scrisă:

$$\sqrt[3]{2 + \frac{1}{9}\sqrt{321}} + \sqrt{2 - \frac{1}{9}\sqrt{321}}$$

utilizați această expresie și calculați cu 4 zecimale corecte pe α.

b) Calculați pe α cu metoda Newton-Raphson tot cu 4 zecimale corecte. plecînd de la $x_0 = 2$.

12. Calculați cu 4 zecimale soluția nenulă a ecuației

$$x = 1 - e^{-2x}.$$

Exerciții facultative (cap. III)

1) Fie $g(x)$ o funcție reală de variabilă reală și $x_0 \in R$. Vom spune că $g(x)$ se comportă bine în x_0 și vom scrie $g \in FB(x_0)$, dacă există două intervale $I_1 = [x_0, x_0 + \varepsilon)$ și $I_2 = (x_0 - \varepsilon, x_0]$, $\varepsilon > 0$ astfel încît $g(x)$ este convexă sau concavă pe I_1 și independent convexă sau concavă pe I_2.

a) Arătați că dacă $g'(x_0) \neq 0$ atunci $g \in FB(x_0)$.

b) Aparține funcția $g(x) = x^2 \sin 1/x$ lui $FB(0)$?

c) Fie $f : R \to R$ cu derivată continuă pe $I = (x_0 - \varepsilon, x_0 + \varepsilon)$ pentru un anumit ε arbitrar. Presupunem că:

$$g(t) = f(x_0 + t) - f(x_0 - t) \text{ are proprietatea } g \in FB(0).$$

Fie:

$$d(h) = \frac{f(x_0 + h) - f(x_0 - h)}{2h},$$

arătați că pentru $h > 0$ suficient de mic $d(h)$ converge monoton către $f'(x_0)$.

d) Pentru $g(t)$ definit la punctul c) cu derivată de ordinul al doilea continuă pe $I = [0, \varepsilon)$, $\varepsilon > 0$ arbitrar și $g, g' \in FB(0)$ arătați că:

$$F(h) = \left| g'(h) - \frac{g(h)}{h} \right|$$

este monoton crescătoare pentru h suficient de mic.

e) În ipotezele de la punctul d) presupunînd în plus că avem un şir de numere pozitive monoton descrescător $\{h_i\}$ care satisface pentru orice $\beta > 1$ relaţia:

$$h_{i+1} = \frac{h_i}{\beta}, \quad i = 1, 2, \ldots$$

arătaţi că:
pentru toţi $i \geqslant N$ (N un anume număr întreg) avem:

$$|d(h_{i+1}) - d(h_i)| \leqslant |d(h_i) - d(h_{i-1})|.$$

f) Tabelaţi derivatele funcţiilor elementare e^x, $\sin x$, $\operatorname{tg} x$, $\sqrt{x}$ pe $[0, 1]$ într-o sută de puncte echidistante utilizînd algoritmul dat de schema logică următoare ($d(h)$ este dat la punctul *c*) schema III. 8.

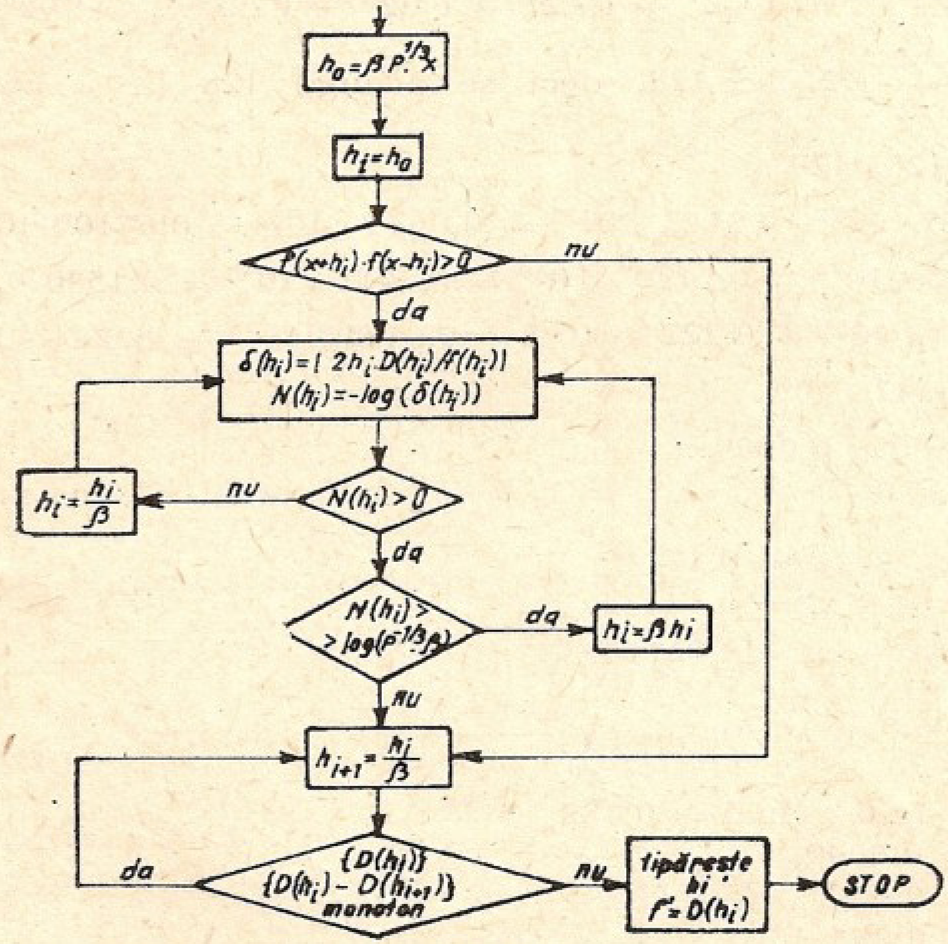


Fig. III. 8

Precizăm că β poate fi luat 4, P este precizia cu care lucraţi, iar dacă $x = 0$ porniţi cu $h_0 = 0{,}01 \cdot 4 \cdot P$.

De asemenea am notat cu:

$$D(H) = (F(X \oplus H) - F(X \ominus H)) \odot (2 \odot H)$$

unde literele mari reprezintă reprezentarea în calcul a literelor mici iar operaţiile încercuite reprezintă aproximarea maşinii la operaţiile aritmetice.

Răspunsuri

Capitolul I

1. 8, 32768, 2660.

2. $a = 0{,}3333 \cdot 10^0$, eroarea $= -1/3 \cdot 10^{-4}$, corecția $= 1/3 \cdot 10^{-4}$.

3. da.

4. numai cînd în reprezentarea $a = m \cdot 10^q + m_1 \cdot 10^{q-t}$ (vezi § 5)

$$0 < |m_1| < 0{,}5.$$

5. $0{,}5 \cdot 10^{-6}$ pentru folosirea rotunjirii și $1 \cdot 10^{-6}$ pentru trunchiere.

6. 43, 56, 28, 39, 80.

7. $1\,1\,1\,1\,1\,1\,1\,1 = 1 \cdot 2^7 + 1 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^4 \ldots + 1 \cdot 2^0 = 2^8 - 1 =$ $= 255;\ \dfrac{1}{2} \cdot 255 \simeq 128$, deci se adaugă 128 la exponentul real $q \in [-128,\ 127]$.

8. *a)* $0{,}2135 \cdot 10^{51} \cdot 0{,}3142 \cdot 10^{51} = 0{,}0671 \cdot 10^{52} = 0{,}6710 \cdot 10^{51}$,

d) $0{,}2135 \cdot 10^{50} - 0{,}2120 \cdot 10^{50} = 0{,}0015 \cdot 10^{50} = 0{,}1500 \cdot 10^{48}$,

e) $0{,}1214 \cdot 10^{51} + 0{,}2221 \cdot 10^{52} = 0{,}0121 \cdot 10^{52} + 0{,}2221 \cdot 10^{52} =$ $= 0{,}2342 \cdot 10^{52}$.

9. 0,087; 0,009; 0,0009 a treia.

10. 0,1.

11. $x_1 + (x_2 + x_3) = 0{,}12115$,
$(x_1 + x_2) + x_3 = 0{,}12114$,
$(x_1 + x_3) + x_2 = 0{,}12114$.

12. *a)* Avem $\quad y_n + 5y_{n-1} = \displaystyle\int_0^1 \frac{x^n + 5x^{n-1}}{x+5}\,\mathrm{d}x = 1/n;$

b) $y_0 = \ln 6 - \ln 5 = 0{,}182$,
$y_1 = 0{,}0090$,
$y_2 = 0{,}050$,
$y_3 = 0{,}083$,
$y_4 = -0{,}165$ (absurd).

c) Eroarea de rotunjire $5 \cdot 10^{-4}$ se amplifică la fiecare pas astfel că obținem în y_4 o eroare de $625 \cdot 5 \cdot 10^{-4} = 0{,}3125$.

13. $y_8 = 0{,}019$, $y_5 = 0{,}028$, $y_2 = 0{,}058$,

$y_7 = 0{,}021$; $y_4 = 0{,}034$, $y_1 = 0{,}088$,

$y_6 = 0{,}025$, $y_3 = 0{,}043$, $y_0 = 0{,}182$ (corect).

Deci avem stabilitate numerică.

Capitolul II

1. Se exprimă ctg α în funcție de parametrii directori ai celor două drepte.

2. Se folosește expresia numărului de condiționare dat de lungimea euclidiană a unei matrice.

3. $\mathrm{cond}_\infty(A) = 10.967$. Soluția sistemului inițial $x = 1$; $y = 1$. Soluția sistemului perturbat $\tilde{x} = -12{,}9$; $\tilde{y} = -20$. Perturbarea soluției $x - \tilde{x} = 13{,}9$; $y - \tilde{y} = 21$.

4. Nu este nevoie.

5. $x_1 = 6{,}154047$, $x_2 = 4{,}313317$, $x_3 = 3{,}240207$.

6. După cinci iterații obținem:

$x_1 = 0{,}49854$,

$x_2 = 0{,}74927$,

$x_3 = 0{,}24927$,

$x_4 = 0{,}49963$.

7. $\mathrm{cond}_\infty\ (A) \geqslant 1{,}5 \cdot \varepsilon^{-1}$.

8. Soluția sistemului inițial $x = -1$; $y = 0$. Soluția sistemului perturbat $\tilde{x} = 1$; $\tilde{y} = 1$.

Se notează $A = \begin{Vmatrix} 1 & -2 \\ -2 & 4{,}01 \end{Vmatrix}$ și se calculează $\|A^{-1}\|_\infty$ care număr se înmulțește cu 0,01 și se deduce că apriori putem avea o modificare de 0,01 în maximum pe componente.

Exerciții facultative. *b*) Se obține relația de recurență

$$\Delta_n = 2\cos\theta - \Delta_{n-1} - \Delta_{n-2}.$$

Capitolul III

3. Un singur pas. Nu.

4. $f(x) = x^3 - 5x^2 - 45x + 50$. Se verifică relația:

$$f(1)\cdot f(2) < 0.$$

5. Aplicînd metoda Newton-Raphson obținem: $x_{n+1} = \dfrac{1}{2}\cdot(2/x_n + x_n)$.

7. $x = 4{,}49341$.

8. Se aplică rezultatele din § 3.

9.

```
SUBROUTINE ITER (FUNC, X0, EPS)
X1 = FUNC(X0)
X2 = FUNC(X1)
```

```
6 X3 = FUNC(X2)
  XM = (X3-X2)/(X2-X1)
  A = XM.(X3-X2)/(1-XM)
  IF(A.LT.EPS) GO TO 7
  X1 = X2
  X2 = X3
  GO TO 6
7 WRITE (108,8) X3
8 FORMAT (1X, E 12.5)
  STOP
  END
```

10. Luați $x_0 = 0{,}317$ și $x_{n+1} = \dfrac{\cos x_n}{3}$ veți obține rădăcina cuprinsă între 0,316749 și 0,316754.

11. 1,7963.

12. Considerăm $f(x) = x - 1 + e^{-2x} = 0$.
Metoda Newton ne dă:

$$x_{n+1} = x_n + h_n;\ \ h_n = -\frac{f(x_n)}{f'(x_n)} = -(x_n - 1 + e^{-2x_n})/(1 - 2e^{-2x_n}).$$

Pentru $x_0 = 0{,}8$ obținem $x_1 = 0{,}79682$ deci $h_1 = -10^{-5}$ deci rădăcina este $\alpha = 0{,}7968$.

Răspunsuri la exercițiile facultative (cap. III)

c) avem: $d(h) = \dfrac{g(h)}{2h}$, prin derivare obținem:

$$(d(h))' = \frac{g'(h)\cdot h - g(h)}{2\cdot h^2} = \frac{g'(h)\cdot h - (g(h) - g(0))}{2\cdot h^2} =$$

$$= \frac{g'(h) - g'(t)}{2\cdot h},\ 0 < t < h \text{ (am aplicat formula lui Lagrange pe } [0, h]).$$

Deoarece $g \in FB(0)$, pentru h suficient de mic $(d(h))'$ are exact un singur semn apoi din $h_1 > h_2$(h_1 și h_2 fac parte din acei h pentru care raționamentul de mai sus merge) deducem:

$$d(h_1) - d(h_2) = \int_{h_2}^{h_1} d(s)ds$$

de unde rezultă monotonia, apoi

$$\lim_{h\to 0} d(h) = f'(x_0) \text{ pentru } f \text{ cu derivată continuă.}$$

d) Deoarece $g \in FB(0)$ putem presupune că g este convexă pentru h suficient de mic, atunci g' este crescătoare pentru h suficient de mic, deci:

$$F(h) = g'(h) - \frac{g(h)}{h}.$$

Prin derivare obținem:

$$F'(h) = g''(h) - \frac{g'(h) - g(h)/h}{h} = g''(h) - \frac{g'(h) - g'(t)}{h};\ 0 < t < h$$

$$= g''(h) - g''(s)(1 - t/h),\ t < s < h.$$

Deoarece $g' \in FB(0)$ urmează că g'' este monotonă. Cum g este presupus convex urmează că $g'' \geqslant 0$ iar $g(t) = f(x_0 + t) - f(x_0 - t)$ ne dă $g''(0) = 0$.

Dacă g'' este monoton crescătoare atunci $F'(h) \geqslant 0$ în timp ce g'' este monoton descrescătoare urmează că $F'(h) = 0$ căci $g'(0) = 0$ și $g'' \geqslant 0$.

e) Urmînd punctul c) putem scrie:

$$|d(h_i) - d(h_{i-1})| = \int_{h_i}^{h_{i-1}} \frac{1}{2h} \left| g'(h) - \frac{g(h)}{h} \right| dh$$

conform lui d) pentru $i \geqslant N$, N un anume întreg pozitiv, avem:

$$|d(h_i) - d(h_{i-1})| \geqslant \frac{|g'(h_i) - g(h_i)/h|}{2} = \int_{h_i}^{h_{i-1}} 1/h \, dh =$$

$$= \ln (h_{i-1}/h_i) \cdot \frac{|g'(h_i) - g(h_i)/h_i|}{2}.$$

Similar avem:

$$|d(h_i) - d(h_{i+1})| = \int_{h_{i+1}}^{h_i} 1/h \left| g'(h) - \frac{g(h)}{h} \right| dh \leqslant$$

$$\leqslant \ln (h_i/h_{i+1}) \cdot \frac{|g'(h_i) - g(h_i)/h_i)|}{2}$$

de unde rezultă concluzia cerută.

Bibliografie

1. Bull Gordon, *Computational Methods and Algol*, George Harop G. & Co LTD., London, Toronto, Wellington, Sydney, 1966.
2. Dahlquist, G. § Ake Björck: *Numerical Methods*, Prentince Hall, INC., 1974,
3. Dorn S., W. § D. D. McCracken: *Metode numerice cu programe în FORTRAN*, Editura tehnică, București, 1976.
4. Ikramov H.: *Recueil de problemes de algebre linéaire*, Edition Mir, Moscou, 1977.

CUPRINS